Addressing treatment contamination in the design and analysis of trials of complex interventions

Magill, Nicholas Peter

Awarding institution:
King's College London

Addressing treatment contamination in the design and analysis of trials of complex interventions

Nicholas Peter Magill

A thesis submitted for the degree of
Doctor of Philosophy

Department of Biostatistics and Health Informatics
Institute of Psychiatry, Psychology and Neuroscience
King's College London
United Kingdom
2018

For Lara

Abstract

In mental health trials there is concern that the control treatment (therapy) might be contaminated. For example, this might be due to it being delivered by a clinician (therapist) who has been trained in the active intervention. It is often suggested by investigators, reviewers and funders of clinical trial proposals that cluster randomisation, with clusters defined at the level at which treatment contamination is thought to occur (e.g. therapists), should be used to prevent contamination. This thesis explores statistical methodologies used to account for both treatment contamination in the control trial arm and non-compliance in the experimental arm when estimating treatment efficacy. It considers trials where treatment receipt is measured on a binary scale and develops methods to accommodate continuous measures of treatment receipt.

The primary research objective was to compare the efficiency of two competing trial design options for evaluating efficacy in the presence of contamination. First, treatment allocation by cluster randomisation together with an estimator of the average treatment effect that accounts for clustered data. Second, allocation at the participant level, measurement of treatment receipt in all participants, and use of a randomisation-based estimator to target the complier average causal effect. Monte Carlo simulations under the two options with a binary measure of treatment receipt showed that the cluster randomisation design was more efficient under high levels of contamination, modest intraclass correlation coefficients and small cluster sizes. With a continuous measure of treatment receipt, the design with cluster randomisation was favoured more frequently as the difference between potential (counterfactual) doses became smaller, i.e. when there was greater non-adherence.

The secondary research objective was to develop a novel randomisation-based efficacy estimator in trials with contamination and non-compliance measured on a continuous scale. Efficacy estimators were applied to a trial of a psychological intervention for people with poorly controlled type 2 diabetes. There was some treatment contamination and non-compliance in the trial but little evidence of treatment efficacy according to the analyses. The tertiary objective was to review problems and solutions associated with contamination in published trials of complex interventions in mental health.

A main output from this project is the proposal of a novel valid estimator for evaluating efficacy and a demonstration of its utility. Another output is the provision of an online decision support tool (using the results from the simulations) to help those planning trials choose between the two competing design options for dealing with contamination.

Acknowledgements

I would like to express my gratitude for the support I have received from many people over the course of this project.

I thank my supervisors, Professor Sabine Landau, Professor Khalida Ismail and Professor Paul McCrone, without whom this work would not have been possible. In particular, I am indebted to Professor Landau for her tireless support, knowledge and enthusiasm.

I am grateful to Professor Ismail for sharing the D6 data with me and to three people who contributed to the fidelity rating of D6 treatment sessions. They are Clare Tucker who listened to and organised the tapes, Pamela Macdonald who conducted the BECCI ratings, and Amy Harrison who performed the MITI ratings. I would like to put on record my thanks to the D6 trial participants for their time and energy.

I am very grateful to the following people from the Department of Biostatistics and Health Informatics for their input and encouragement: Deborah Agbedjro, Cedric Ginestet, Dominic Stringer, John Hodsoll, Ruijie Li and Jennifer Hellier. I would also like to express my appreciation to Ruth Knight for checking some of the study selection and data extraction in the scoping review. I am grateful to Meera Ladwa for valuable insight into the treatment of diabetes.

I thank my family for their support, in particular my dad Patrick for comments on the thesis and my wife Jules for always being there.

Contents

List of figures **8**

List of tables **10**

List of abbreviations **12**

1 General introduction **14**
1.1 Randomised controlled trials . . . 14
1.2 Treatment contamination in trials of complex interventions in mental health 15
1.3 The problem of contamination . . . 16
1.4 The causes, extent and impact of contamination . . . 17
1.5 Methods of addressing contamination . . . 19
1.5.1 Statistical design methods . . . 20
1.5.2 Trial conduct methods . . . 24
1.5.3 Analytical methods . . . 25
1.6 D6 trial . . . 32
1.6.1 Background . . . 32
1.6.2 Trial design . . . 34
1.6.3 Setting and target population . . . 34
1.6.4 Active intervention . . . 35
1.6.5 Comparator treatment . . . 35
1.6.6 Treatment allocation . . . 35
1.6.7 Outcomes . . . 36
1.6.8 Sample size . . . 36
1.6.9 Evaluation of clinical effectiveness . . . 37
1.6.10 Treatment contamination . . . 37
1.7 Other trials motivating this research . . . 38
1.8 Aims of the thesis . . . 39
1.9 Thesis overview . . . 40

2 Scoping review of problems and solutions associated with contamination **43**
2.1 Background and aims . . . 43
2.2 Methods . . . 44
2.2.1 Type of review . . . 44
2.2.2 Eligibility criteria . . . 45
2.2.3 Information sources . . . 45
2.2.4 Search . . . 45
2.2.5 Study selection . . . 46
2.2.6 Data collection process . . . 46
2.2.7 Data items . . . 47

2.2.8 Risk of bias in individual studies . . . 47
2.3 Results . . . 47
2.3.1 Reliability . . . 47
2.3.2 Summary of trials . . . 47
2.3.3 Summary of assessment of bias . . . 48
2.3.4 Processes driving contamination . . . 50
2.3.5 Quantity of contamination . . . 52
2.3.6 Solutions used to counter contamination . . . 56
2.3.7 Trials using both cluster- and participant-level treatment allocation 58
2.3.8 Parameters . . . 59
2.4 Discussion and conclusion . . . 62
2.4.1 Discussion . . . 62
2.4.2 Conclusion . . . 64

3 Assessing treatment fidelity and contamination in a cluster randomised controlled trial of motivational interviewing and cognitive behavioural therapy skills in type 2 diabetes. 65
3.1 Background and aims . . . 65
3.2 Methods . . . 67
3.2.1 The training programme . . . 67
3.2.2 Techniques taught in MI and CBT . . . 68
3.2.3 Clinical supervision . . . 68
3.2.4 Assessment of treatment fidelity and competency . . . 68
3.2.5 Nurse competency assessment . . . 70
3.2.6 Sampling for inter-rater reliability assessment . . . 70
3.2.7 Sampling for fidelity assessment . . . 71
3.2.8 Statistical analysis . . . 71
3.3 Results . . . 72
3.3.1 Nurse and patient sample characteristics . . . 72
3.3.2 Nurse competency . . . 73
3.3.3 Inter-rater reliability . . . 73
3.3.4 Fidelity analysis . . . 74
3.4 Discussion and conclusion . . . 75
3.4.1 Discussion . . . 75
3.4.2 Conclusion . . . 78

4 Efficacy estimators allowing for non-adherence in trials of complex interventions 79
4.1 Background and aims . . . 79
4.2 Estimands . . . 80
4.2.1 Rubin causal model . . . 80
4.2.2 Potential outcomes . . . 81
4.2.3 Identification assumptions of Rubin causal model . . . 84
4.3 Traditional estimation approaches . . . 85
4.3.1 Estimating effectiveness: intention-to-treat estimator . . . 85
4.3.2 Estimating efficacy: as treated and per protocol . . . 86
4.4 Randomisation-based efficacy estimators for addressing non-compliance . 87
4.4.1 Principal stratification . . . 88
4.4.2 Instrumental variables . . . 92

4.5 Randomisation-based efficacy estimation approaches for addressing contamination . . . 106
4.5.1 Principal stratification . . . 106
4.5.2 Instrumental variables . . . 109
4.6 Discussion . . . 116

5 Monte Carlo simulation study comparing two trial design options for addressing contamination 119
5.1 Background and aims . . . 119
5.2 Simulation study 1: Binary measure of treatment receipt under design option B . . . 121
5.2.1 Contamination process in therapy trial . . . 121
5.2.2 Data generating models . . . 123
5.2.3 Simulation study design . . . 130
5.2.4 Simulation findings . . . 134
5.3 Simulation study 2: Continuous measure of treatment receipt under design option B . . . 142
5.3.1 Contamination process in therapy trial . . . 142
5.3.2 Data generating models . . . 144
5.3.3 Simulation study design . . . 155
5.3.4 Simulation findings . . . 159
5.4 Discussion . . . 171

6 Decision support tool 173
6.1 Background and aims . . . 173
6.2 Graphical demonstration of simulation results . . . 175
6.3 Online decision support tool . . . 176
6.3.1 Specification and design . . . 176
6.3.2 Graph toggles . . . 178
6.3.3 Online publication . . . 179
6.3.4 User instructions . . . 179
6.4 Demonstration of the decision support tool . . . 182
6.4.1 Planning a trial with a binary measure of treatment receipt . . . 182
6.4.2 Planning a trial with a continuous measure of treatment receipt (dose) . . . 187
6.5 Conclusion . . . 188

7 Estimating treatment efficacy under treatment contamination: Application to the D6 trial 190
7.1 Background . . . 190
7.2 Aims and hypotheses . . . 191
7.3 Method . . . 192
7.3.1 Statistical issues . . . 192
7.3.2 Generation of adherence measures . . . 192
7.3.3 Predictors of treatment receipt and missingness . . . 195
7.3.4 Efficacy estimators . . . 195
7.3.5 Software implementation . . . 196
7.4 Results . . . 200
7.4.1 Patient sample characteristics . . . 200
7.4.2 Adherence with allocated treatments . . . 200

7.4.3 Missingness of HbA_{1c} and predictors of it . . . 203
7.4.4 Effectiveness assessment . . . 203
7.4.5 Efficacy assessment . . . 207
7.5 Discussion . . . 210

8 General discussion **215**
8.1 Overview of thesis . . . 215
8.2 Main findings . . . 216
8.2.1 Primary research objective . . . 216
8.2.2 Secondary research objective . . . 218
8.2.3 Tertiary research objective . . . 219
8.3 Implications for trials . . . 220
8.3.1 Statistical design . . . 220
8.3.2 Conduct . . . 223
8.3.3 Analysis . . . 224
8.3.4 Reporting . . . 225
8.4 Contribution to statistical methods for dealing with contamination and non-compliance . . . 226
8.5 Contribution to the field of diabetes treatment in the context of psychological medicine . . . 228
8.6 Strengths and limitations . . . 230
8.7 Future work . . . 233

References **235**

Appendix A Scoping review **254**
A.1 Ovid search procedure . . . 254
A.2 Pro forma for information/data extraction . . . 257
A.3 Articles from which data were abstracted in the scoping review . . . 260

Appendix B Shiny application **283**
B.1 "ui.R" script . . . 283
B.2 "server.R" script . . . 291
B.3 Print output script . . . 296
B.4 3D plot script . . . 296

Appendix C Application of efficacy estimators to D6 **299**
C.1 Stata code for estimator E-IV6 (Bloom/ratio) with bootstrap standard error 299
C.2 Stata code for estimator E-IV5 (modified Bloom/ratio) with bootstrap standard error . . . 300
C.3 MPlus code for STR3 estimator . . . 300

List of figures

1.1 Illustration of typical randomised controlled trial demonstrating the target effect of effectiveness. . . . 26
1.2 Illustration of typical randomised controlled trial with non-adherence and demonstrating the target effect of efficacy. . . . 27

2.1 Flow diagram for searching for relevant articles . . . 48
2.2 Processes driving treatment contamination in trials. . . . 55
2.3 Forest plots for four trials that used both individual- and cluster-level randomisation. . . . 61

4.1 Structural equation model diagram illustrating the causal effect of intervention allocation on outcome. . . . 86
4.2 Structural equation model diagram illustrating use of principal stratification to address the problem of treatment non-compliance. . . . 90
4.3 Structural equation model diagram illustrating use of principal stratification to address the problem of treatment non-compliance, valid under latent ignorability. . . . 92
4.4 Structural equation model diagram assumed to estimate the causal effect of treatment on outcome. . . . 93
4.5 Structural equation model diagram illustrating use of principal stratification to address the problem of treatment contamination. . . . 107
4.6 Structural equation model diagram illustrating use of principal stratification to address the problems of treatment contamination and non-compliance. . . . 108

5.1 Distributions of $D(0)$ and $D(1)$ for the simulations of populations with full dose compliers and strong partial dose compliers. . . . 160
5.2 Distributions of $D(0)$ and $D(1)$ when simulating populations with moderate strength and weak partial dose compliers. . . . 161

6.1 Three-dimensional plot of the isosurface representing equivalence of estimator standard errors between design options for a binary measure of treatment receipt. . . . 177
6.2 Three-dimensional plot of the wireframe box with arrows representing changes in viewing angle for two of the toggles. . . . 179
6.3 Application of the decision support tool to a trial based on the D6 study and defining treatment receipt as binary. . . . 185
6.4 Application of the decision support tool to a hypothetical trial with binary treatment receipt and plausible levels of sample size, ICC, cluster size and proportion of non-contaminators. . . . 186

6.5 Application of the decision support tool to a hypothetical trial with continuous treatment receipt and plausible levels of sample size, ICC, cluster size, proportion of dose compliers and magnitude of dose compliance within this stratum. 189

7.1 Plot of means and 95% confidence intervals of standard care and D6 intervention arms over time for the primary analysis (full trial) sample. . 206
7.2 Plot of estimated causal treatment effects associated with maximal difference in MITI Global Spirit between the counterfactual situations at all post-randomisation time points. 211

8.1 Flowchart demonstrating the more efficient trial design method for the estimation of either effectiveness or efficacy in trials with contamination. 222

List of tables

1.1 Mean scores of treatment fidelity scales (MITI and BECCI) by treatment group from D6's primary assessment. 38
1.2 Sources of data in the PhD and uses of the datasets in this research. . . . 41

2.1 Summary of characteristics of trials (n=238 trials). 49
2.2 Summary of assessment of bias in trials. 50
2.3 Quantity of treatment contamination in trials where participants could either receive or not receive treatment (i.e. treatment receipt measured on binary scale). 52
2.4 Trial conduct solutions to treatment contamination. 57
2.5 Summary of trials using both cluster- and participant-level treatment allocation. 58
2.6 Summary of clustered data and sample size parameters. 59

3.1 Minima, maxima, and thresholds for MITI and BECCI scales. 69
3.2 Summary of competency scores assessed after training. 73
3.3 Inter-rater reliability for MITI global scores and BECCI Practitioner Score. 73
3.4 MITI summary scores during treatment delivery by treatment allocation group. 74
3.5 Numbers and proportions of sessions rated as above MITI's "Beginning proficiency" and "Competency" thresholds for domains by treatment allocation group. 75

4.1 Notation for observed variables. 81
4.2 Principal strata. 83
4.3 Observed compliance status. 88
4.4 Grid of ACE_{d_1,d_0} at levels of dose when offered control (d_0) or active intervention (d_1). 114

5.1 Summary of input levels of simulation parameters when simulating a binary measure of treatment receipt. 130
5.2 Absolute bias of ITT estimator (option A) for binary treatment receipt (no treatment non-compliance). 134
5.3 Absolute bias of as-treated estimator (option B1) for binary treatment receipt (no treatment non-compliance). 135
5.4 Absolute bias of IV estimator (option B2) for binary treatment receipt (no treatment non-compliance). 136
5.5 Model SE / true SE of ITT estimator (option A) for binary treatment receipt (no treatment non-compliance). 137

5.6 Model SE / true SE of IV estimator (option B2) for binary treatment receipt (no treatment non-compliance). 137
5.7 Monte Carlo standard error for ITT estimator (option A) for binary treatment receipt (no treatment non-compliance) 138
5.8 Monte Carlo standard error for IV estimator (option B2) for binary treatment receipt (no treatment non-compliance) 138
5.9 Relative efficiency of design options A and B2, with binary treatment receipt and no non-compliance in the intervention arm. 140
5.10 Relative efficiency of design options A and B2, with binary treatment receipt and 20% non-compliance in the intervention arm. 141
5.11 Summary of input levels of simulation parameters when simulating a continuous measure of treatment receipt. 156
5.12 Absolute bias of as-treated estimator (option B1) for continuous treatment receipt (when dose responders were full compliers). 162
5.13 Absolute bias of IV estimator (option B2) for continuous treatment receipt (when dose responders were full compliers). 163
5.14 Model SE / true SE of IV estimator (option B2) for continuous treatment receipt (when dose responders were full compliers). 164
5.15 Relative efficiency of design options A and B2, with continuous treatment receipt (when dose responders were full dose responders). 167
5.16 Relative efficiency of design options A and B2, with continuous treatment receipt (with strong partial dose responders). 168
5.17 Relative efficiency of design options A and B2, with continuous treatment receipt (with moderate strength partial dose responders). 169
5.18 Relative efficiency of design options A and B2, with continuous treatment receipt (with weak partial dose responders). 170

7.1 MITI and BECCI domains with definitions. 193
7.2 List of estimators from Chapter 4 that were applied to D6 primary outcome data. 196
7.3 Summary of patient characteristics for the treatment fidelity assessment sample. 201
7.4 Unadjusted odds ratios for possible predictors of treatment receipt. 202
7.5 Unadjusted odds ratios for possible predictors of glycated haemoglobin (HbA_{1c}) response at 18 months after randomisation. 204
7.6 Summary of HbA_{1c} (mmol/mol) at outcome time points for both the full trial and treatment fidelity assessment samples. 205
7.7 Estimated difference in HbA_{1c} (mmol/mol) between treatment groups at outcome time points using effectiveness estimator (ITT analysis). 205
7.8 Estimated differences in HbA_{1c} (mmol/mol) at outcome time points using efficacy estimators with binary measure of treatment receipt. 208
7.9 Estimated differences in HbA_{1c} (mmol/mol) at outcome time points using efficacy estimators with continuous measure of treatment receipt. 211

List of abbreviations

2SLS two stage least squares.

ACE average causal effect.

ATE average treatment effect.

ATR adjusted treatment response.

ATT average treatment effect on the treated.

BECCI Behavioural Change Counselling Index.

BMI body mass index.

CACE complier average causal effect.

CBT cognitive behavioural therapy.

CRCT cluster randomised controlled trial.

D6 diabetes-6.

FIML full information maximum likelihood.

G2SLS generalised two stage least squares.

HbA_{1c} glycated haemoglobin.

ICC intraclass correlation coefficient.

IRE individual randomisation effect.

ITT intention-to-treat.

IV instrumental variable.

LATE local average treatment effect.

LI latent ignorability.

LIML limited information maximum likelihood.

MAR missing at random.

MCAR missing completely at random.

MCSE Monte Carlo standard error.

MI motivational interviewing.

MITI Motivational Interviewing Treatment Integrity.

ML maximum likelihood.

OLS ordinary least squares.

RCT randomised controlled trial.

SEM structural equation model.

SUTVA stable unit treatment value assumption.

T2D type 2 diabetes.

TAU treatment as usual.

Chapter 1

General introduction

1.1 Randomised controlled trials

Randomised controlled trials (RCTs) are a mainstay of quantitative research and are often described as the "gold standard" of study design. This is because, of all study designs, they provide the most convincing evidence of the effect of a treatment. The evidence that a well-conducted RCT provides on treatment effect can be described as internally valid because such a study is capable of giving an answer to a research question that is unbiased. This is a consequence of treatment allocation being random and therefore not driven by other factors that affect both the delivery of the exposure and the outcome. This confers a causal status on the evidence that RCTs provide about the effect of an intervention. However, trials are often expensive, take years to deliver and, particularly in mental health, struggle to recruit enough participants (Campbell et al., 2007; Howard et al., 2009). This implies that it is important that clinical trials are designed to be effective and efficient. In other words they should provide valid answers to research questions and make best use of resources and participants' time and data.

The ideal RCT would recruit a representative sample from a population, allocate treatment randomly, conceal this allocation from participants, clinicians and outcome assessors, have no loss of research participants before outcomes are collected, and feature full treatment adherence with protocol. Therefore, in an ideal RCT, the only difference between trial arms (the groups of participants who are allocated to different interventions) would be the treatment that participants are allotted to receive. In reality, trials can experience problems with all of these ideal characteristics. It is one aspect of the last in the list that this research project will focus on: treatment contamination.

1.2 Treatment contamination in trials of complex interventions in mental health

The general definition of treatment contamination in RCTs is the receipt of a treatment that has been allocated to at least one trial arm by participants in another arm. This could involve the receipt of the comparator treatment by participants in an experimental arm, the receipt of an experimental treatment by control participants, or the receipt of the experimental treatment by participants in another experimental trial arm. A narrower and more common definition of contamination is the "process whereby an intervention intended for members of the trial (intervention or treatment) arm of a study is received by members of another (control) arm" (p. 6, Keogh-Brown et al., 2007). This thesis will use this definition of treatment contamination throughout. There are a number of synonymous terms for contamination. It is known occasionally as "spillover" (Ell et al., 2010), "crossover" (Abroug et al., 2011), "diffusion" (Taylor et al., 2005), or "intrusion" (Pearl, 2009).

In an RCT, treatment contamination is a type of treatment (or protocol) non-adherence. Non-adherence can be considered separately amongst those allocated to active intervention compared to those assigned to control. Amongst those in the former group, the problem of participants not receiving the active intervention is known as non-compliance. On the other hand, non-adherence in the control group is known as treatment contamination. The problems of non-compliance and contamination are related but may have very different solutions (for example design methods for preventing contamination that will be described in Section 1.5.1 would not be used to address non-compliance). Trialists tend to be more accepting of non-compliance, partly because there is restricted scope for preventing it by trial design but also because its effects are limited to the individual. Contamination, on the other hand, causes more concern due to the possibility of it spreading through a group of participants quickly and widely in extent, and because it can be harder to measure (e.g. it may be difficult to record what treatments participants have received outside a trial). On the other hand, as will be described, there is a similarity in terms of the analytical methods that can be used to address the two problems (see Section 1.5.3).

Complex interventions, which are defined as packages of medical care with multiple elements that produce some extra benefit when given together, are a mainstay of tre-

atment in mental health and psychiatric disease (Craig et al., 2008). However, they can be especially vulnerable to treatment contamination because their components are sometimes "transportable and difficult to confine" (quotation used in the description of another type of complex intervention in health research: educational interventions; p. ix, Keogh-Brown et al., 2007).

1.3 The problem of contamination

The primary research question of the majority of late-phase trials is concerned with effectiveness and this is often estimated using an intention-to-treat (ITT) analysis, where treatment groups are defined by which trial arm participants were randomly allocated to (more on this in Section 1.5.3). The effect of contamination is to dilute the difference between trial arms, as the control group becomes more similar to the treatment group. It is frequently considered that its consequence is to reduce the magnitude of the treatment effect estimate and increase the size of this parameter's estimated standard error (e.g. Torgerson, 2001; Welsh, 2013). In the context of a superiority trial where the target effect is effectiveness, the impact of both of these processes is to decrease statistical power and therefore reduce the chance of detecting a statistically significant effect.

An ITT analysis in the context of contamination is often referred to as being "conservative" due to the fact that the chance of detecting a statistically significant effect is reduced. It should be emphasised that the impact of contamination is only conservative in the context of a superiority trial where the treatment shows some evidence of benefit. Its effect would be to give false evidence of effect if the trial were a non-inferiority (equivalence) design or of safety if the treatment were harmful. It should also be considered that whilst a bias towards the null may be reassuring to those in charge of organising a healthcare system, it is less informative to patients and clinicians who may want to know the effect of treatment in ideal conditions where the effect is likely to be greatest.

The problem of contamination is seemingly related to lack of blinding in trials. Based on the focus of trial design and analysis methods in the existing trials literature, the phenomenon appears to be of concern in educational interventions (Keogh-Brown et al., 2007), geriatric medicine (Borm et al., 2005; Melis et al., 2008), cancer screening (Cuzick et al., 1997), and mental health (Dunn et al., 2005). It tends to be difficult or impossible to keep research participants and clinicians blind to treatment allocation in these areas

of research. This is because the content of an intervention in these areas is often obvious to the clinician as they give the treatment and to the patient as they receive it. It is implicitly assumed by researchers that the presence of such knowledge can somehow induce clinician or participant behaviour that leads to those in the control arm of a trial receiving the active intervention.

1.4 The causes, extent and impact of contamination

The trials methodology literature is not entirely clear about the possible processes that lead to contamination. Many RCTs have described these processes, for example this could be a clinician providing both active and control interventions who mixes up elements of the treatments or could be contact between participants in different trial arms leading to a dilution of the treatment contrast (Borm et al., 2005). However, these processes, their relative frequencies and their impact on trial arm comparisons have never been reviewed comprehensively. An understanding of this is important in order to plan what steps should be taken to address the problem.

It has been argued that the mechanism of contamination depends on the level at which an intervention is applied: patient, clinician, or member of the general public (Keogh-Brown et al., 2007). That research envisaged, with particular reference to trials of educational interventions, that patient-level contamination affects only the individual and is less likely for more complex interventions whose ingredients are less transportable. At clinician level, contamination is likely to occur through communication between either active intervention clinicians and control participants, or between clinicians in different trial arms. At the population level, contamination might occur by unintended delivery of intervention throughout the population, e.g. a public health intervention broadcast to control participants. By surveying trialists, the same study showed that contamination was more likely in settings where participants interacted closely, where the intervention was highly desirable, or where the treatment was aimed at increasing knowledge. They also found that contamination was thought to be more likely for interventions applied at clinician rather than patient level.

Previous reviews have assessed the extent of contamination in some areas of medicine, but the scale of the problem in trials of complex interventions in mental health remains unclear. For example, a literature review of 235 RCTs of guideline dissemination and

implementation strategies for healthcare professionals identified eight trials that quantified contamination (Keogh-Brown et al., 2007). The review reported the proportion of patients in the control arm who were assessed as having received treatment and found a median of 24% of patients to be contaminated (range 0-65%). In oncology, a large breast cancer screening trial (n=9,780) found that 22% of those in control arm received a mammogram outside the trial compared to 5% of the intervention group doing likewise (Goel et al., 1998). It has been argued that a review of cancer trials using Zelen's design provided information on treatment switches that was analogous to contamination (Torgerson, 2001; Altman et al., 1995). Trials that use Zelen's method ask patients for consent to allocated treatment before asking for consent to participate in the study. This means that the proportion of controls not giving consent at stage one may provide some information on treatment contamination. The review found 11 trials that reported treatment switches with an average of 18% (range 10-36%). However, many of the studies in the review described switches from active to comparator treatments or provided an overall summary of switches in either direction. Given the reported information, the analogy to contamination was tenuous.

A number of studies have investigated the link between contamination (or the prevention of it) and the size of estimated treatment effect. Many such studies assume that the impact of contamination is to blur the distinction between trial arms and therefore hypothesise that effects will be larger in trials that avoid contamination compared to those that do not. Research that has tested this prediction has shown mixed results. For example, the report that investigated contamination in educational intervention RCTs, which was described earlier (Keogh-Brown et al., 2007), re-assessed an earlier systematic review of educational, financial, and strategic interventions (Grimshaw et al., 2004). In the original review the authors had assessed for each trial whether contamination avoidance had been done, not done, or was not clear. Trials that avoided contamination were usually those that used cluster randomisation. Surprisingly, the results of the later review showed that trials where contamination was avoided had smaller estimated effect sizes than did those where contamination avoidance was not done or not clear. The weakness of this finding was that the studies represented a wide range of study quality and substantial heterogeneity of experimental intervention. When the sample was restricted to those trials that were rated as being of higher quality, estimated treatment effects were again found to be greater for trials where contamination was not avoided. The set was then further restricted to trials that were both of higher quality and in addition tested similar

interventions. These were studies of clinician reminder interventions and provided a sample of 11 trials. The authors now compared estimated treatment effect sizes between cluster and individual randomised trials. Average estimated effect size amongst the cluster randomised controlled trials (CRCTs) avoiding contamination was higher than the average for the individual randomised trials with possible contamination, implying that contamination avoidance may have been successful.

Reviews of particular interventions have found similarly mixed results. For example, a review of 14 hip protector trials showed large beneficial effects of treatment in CRCTs aimed at avoiding contamination and a mixture of positive and negative effects in RCTs with patient-level randomisation with suspected contamination (Hahn et al., 2005). The authors speculated that this difference may have been due to contamination biasing results in the individual randomised trials. Other explanations were that the settings may have been different between the two types of trials or that the results of the CRCTs may have been biased by problems associated with cluster randomisation (see Section 1.5.1). A meta-analysis of 14 individual randomised and 10 cluster randomised trials of enhanced care for depression provided rather different evidence of the relationship between level of treatment allocation and estimated effect size (Gilbody et al., 2008). The review found that estimated mean differences were very similar between groups of trials as defined by level of treatment allocation. The only difference was that individual randomised trials demonstrated substantially more heterogeneity between effect sizes than did the CRCTs.

In summary, it is difficult to draw a conclusion about the relationship between the presence of contamination and size of estimated treatment effects from the available evidence. This is partly because of the heterogeneous nature of the interventions being compared and also because it is difficult to disentangle the impacts of bias due to contamination and those biases associated with the use of cluster randomisation. The evidence would be more persuasive if treatment effect estimates were compared between studies where the only difference was the level of treatment allocation, for example study designs that make use of both cluster randomisation to avoid contamination and individual randomisation.

1.5 Methods of addressing contamination

I distinguish between three types of methods that are used to address contamination. These categories are named here as statistical design, trial conduct and analytical met-

hods. The first refers to structural aspects of trial design such as the level at which treatment is allocated, sample size inflation, and participant preference designs. The second includes trial implementation methods that are used to minimise the process by which participants' treatment becomes contaminated. The third relies on there being a measurement of individual treatment receipt and includes comparison of groups based on treatment received using randomisation-based estimation methods from the causal inference literature.

1.5.1 Statistical design methods

Cluster randomisation

The most common statistical design method of avoiding contamination is the use of cluster randomisation, where groups of participants instead of individuals are allocated to trial arms. This strategy is often advocated by researchers and funders because it can prevent contamination, provided that treatment allocation is made at the highest level at which it is thought to take place (Campbell and Grimshaw, 1998). By ensuring that all participants within a cluster receive the same treatment, contamination of the control condition due to participants being affected by each other's treatment receipt can be avoided.

Cluster randomisation is used frequently in mental health trials, partly because of the problem of contamination, but also for logistical reasons. For example, cluster randomisation can reduce the number of clinicians who need to be trained in a new treatment in comparison to an individually-randomised trial. Cluster randomisation can also lead to a feeling of fairness within communities, as all participants are allocated to the same treatment. However, the use of cluster randomisation cannot always minimise contamination. There is at least one type where the scale of the problem is independent of the level at which treatment is allocated: control participants seeking out the active intervention outside the trial. This is a particular problem in screening trials which are typically found in other areas of medicine, for example in oncology (Pinsky et al., 2010). It is also possible that this problem may be becoming greater as patients are encouraged to take greater responsibility for disease management decisions and with the growth of online forums.

There are substantial drawbacks of cluster randomisation in terms of estimator efficiency and bias. The main cost is that the correlation between participants' outcomes within

clusters means that each additional observation provides less information than it would in an individually-randomised trial. This correlation and the number of participants per cluster must be factored into a power calculation and will inflate the sample size requirement. The design factor (D) that is used for this inflation is:

$$D := 1 + I(k-1)$$

where I is the intraclass correlation coefficient (ICC) and k is the number of members per cluster. These two factors can easily lead to a large inflation of the required sample size. For example an ICC of 0.05 and a cluster size of 20 would almost double the sample size.

The other weakness of CRCTs is increased risk of bias. They have been found to suffer from three main types: selection bias, attrition bias, and assessment bias. The first of these is a consequence of the fact that it is often difficult to recruit all participants to a cluster before allocating a treatment to this group, meaning that allocation concealment is compromised. A review of 36 CRCTs that were published in three prominent journals between 1997-2002 found that only 14 trials identified participants before treatment allocation took place; this represented 41% of the sample where this item could be clearly assessed (Puffer et al., 2003). The implication of this is that the RCT may suffer from trial arm imbalance in terms of patient characteristics. In addition to this, it is possible for whole clusters to be lost to follow-up in CRCTs (attrition bias) and it becomes harder to blind outcome assessors to participants' treatment allocation than it is when allocation is made at the level of the individual (assessment bias; Puffer et al., 2003).

The major report of contamination in educational interventions compared bias in cluster and individual randomised trials using Monte Carlo simulations (Keogh-Brown et al., 2007). They investigated the impact of contamination on the bias of effectiveness estimates (presumed to be the bias of the ITT approach for the effect of treatment receipt on outcome) where cluster randomisation was modelled not to prevent receipt of intervention in the control arm. In the simulations the authors modelled three aspects of contamination: the proportion of the control group that were exposed to the active (educational) intervention, the intensity of this exposure, and in CRCTs the timing of control individuals' exposure to intervention in comparison to when the cluster was first contaminated. They simulated levels of sample size, ICCs, cluster size and timing of contamination (using different levels of Weibull distribution parameters). A parameter was used to relate baseline education level to risk of contamination. The outcome of interest

in the simulations was bias: the difference between the true treatment effect and the estimated effect of intervention offer. The findings suggested that cluster randomisation led to greater bias (than allocation of treatment to individuals) when contamination of a control cluster resulted in all individuals within that group receiving the treatment, and when contamination effects were constant over time. CRCTs were found to be more biased when the interval between the start of the trial and clusters being contaminated was short. When there was a strong link between control participants' baseline need for the intervention and the chance of receiving the active intervention, CRCTs were more biased when the transferability of the intervention between the trial arms was low.

Sample size inflation of individual randomised trial

A simple way of addressing the problem of reduced statistical power caused by contamination in a superiority trial, when the target treatment effect of interest is the effect of treatment offer (effectiveness), is to inflate the required sample size (Torgerson, 2001). Under this method, treatment allocation is applied at the level of the individual and the proportion of the control group who are anticipated to receive the active intervention is used to inflate the sample size, thereby recovering the lost power. This method was compared to a typical CRCT, which was assumed by the study to double the required sample size and to prevent any contamination. On this basis, the sample size of an inflated individually-randomised trial was less than that of a CRCT when contamination was less than 30%. A similar study compared sample size inflation to adjust for contamination in an individually-randomised trial with cluster randomisation (Slymen and Hovell, 1997). The study simulated a range of amounts of contamination, cluster sizes, and ICCs. It found that cluster randomisation led to smaller sample sizes when cluster size and ICC were low, and when the amount of contamination was high. The limitation of this method of sample size inflation is that it does not address the effect of contamination bias on the magnitude of the estimated effect of treatment receipt on outcome.

Treatment allocation at a mixture of cluster and individual levels

A class of methods that balance the benefits of both cluster and individual randomised trials and can be used to address contamination are those that apply treatment allocation at a mixture of levels. For example, pseudo cluster randomised trials with two trial arms randomly split participants into two portions of clusters. In the first of these, the majority

of participants within a cluster are randomised to one treatment and a minority to the other (often using a 80:20 ratio); in the remaining clusters allocation is *vice versa* (Borm et al., 2005). This approach is efficient because it reduces the magnitude of clustering in comparison to a CRCT (not all participants within a cluster receive the same treatment), and it lessens the possibility of selection bias because an individual's treatment allocation is unknown at study entry. It can also reduce the chance of contamination in the specific circumstance where this is caused by control clinicians learning elements of the treatment from active intervention participants and then passing these onto control participants (Teerenstra et al., 2006). Another approach of the same class is to allocate treatment at the cluster level for a subsample of participants and at the individual level for the remainder of the trial sample. This method limits contamination (amongst those who are cluster randomised) and a comparison of treatment effect estimates between the cluster- and individual-randomised subtrials may provide some information as to the impact of contamination or the ability of cluster randomisation to prevent it.

Zelen's and participant preference designs

Two other statistical designs can be used to reduce contamination, although they are rarely used purely for this purpose. The first randomly allocates treatment to participants before asking for consent to take part in the trial (Zelen, 1979). In a double consent Zelen's design trial, participants are initially offered the treatment that they were randomly allocated to and are asked if they consent to receive the treatment. Those who decline the active intervention are given the control treatment and those who refuse the control treatment are offered the active intervention. This method can lessen contamination later on in the trial by reducing feelings of disappointment amongst participants who are allocated to their non-preferred treatment. In addition, Zelen's method is useful for measuring the proportion of the control group whose treatment may be contamination (as described earlier) and could also be useful for identifying these participants' characteristics.

The second, minor statistical design for reducing contamination is the participant preference design. Trials using this design give a proportion of participants a choice about which treatment they receive and then allocate treatment randomly for the rest (Floyd and Moyer, 2010). The article described participants' preferences in an unblinded randomised trial using a doubly randomised participant preference design. This strategy initially randomises participants to either a preference controlled trial or a randomised

trial (Wennberg et al., 1993). Those allocated to the former receive their preferred treatment; those allocated to the second are randomised again and offered whatever treatment is the outcome of this second allocation. The trial intervention was classical music offered to students who were preparing for a college examination. The aim of the research was to assess whether those who were allocated to treatment by choice would show lower contamination compared to those allocated to treatment by chance. Unfortunately, the amount of contamination was too low for this to be assessed (this could also be interpreted as evidence that allocation by choice or chance did not affect contamination). The trial showed that whilst those who were allotted to their non-preferred treatment felt less positive about their experience, this did not affect belief in effectiveness of treatment. Like Zelen's design, participant preference design trials can enable a better understanding of participants' choice and the impact of this on treatment effect. They may limit contamination due to feelings of demoralisation, but this is likely to be effective only when the proportion of the control group who wish to receive the treatment is small.

1.5.2 Trial conduct methods

Very little methodological research appears to have addressed the trial conduct or implementation methods that trialists use to minimise contamination. The only study I am aware of that has attempted to assess these methods was a survey that was conducted as part of the major report on trials of educational interventions (Keogh-Brown et al., 2007). The authors designed a questionnaire that was sent to 100 experts in educational health research. This group included researchers known by the authors, those who were active in the Cochrane Collaboration and Campbell Collaboration, and members of the Association for the Study of Medical Education. Thirty-seven people responded in the first round. The results were summarised and then sent back to the experts who were then invited to rank responses as to which trial designs might lead to the highest chance of contamination. Twenty-seven (73%) replied in the second round.

The results provided some information on which trial conduct methods may be most effective at minimising the chances of contamination. It was thought to be least likely in trials of interventions where their elements were difficult to transfer or where they involved attendance of a training programme or event. The chance of contamination was considered low in settings where participants were unlikely to share social networks and where they were geographically separated. Finally, the use of education of participants

against transfer of the intervention to the control arm and provision of clear information on the purpose of the study were thought to reduce the chance of contamination.

In summary, little research has attempted to review trial conduct or implementation methods for minimising contamination. The only study I am aware of sought the opinions of expert trialists. There will therefore be some merit in reviewing the trials literature and assessing the methods that are used in practice. This knowledge could be used to provide those designing and funding trials where contamination is a concern with methods for minimising the problem without the previously discussed and substantial costs of cluster randomisation.

1.5.3 Analytical methods

Target treatment effects

As described earlier, the primary analysis of a randomised trial often uses the ITT approach. The ITT approach is typically used as a method of estimating effectiveness, which is considered to be the effect of treatment offer on outcome (Hernán and Hernández-Díaz, 2012). This target treatment effect is sometimes known as the *de facto* effect (Mallinckrodt et al., 2017). It is of particular interest to those designing and planning the delivery of health services because it describes the impact of a treatment in real-world conditions. The ITT approach is described as "conservative" because estimates are biased towards the null (only in the context of a superiority trial; Hernán and Hernández-Díaz, 2012). The approach is linked to the design of a pragmatic trial. RCTs that are described as such take place in a 'naturalistic' or 'real-world' clinical setting and have few exclusion restrictions – this is in order to maximise external validity, thereby making the findings generalisable to a wide population (Revicki and Frank, 1999; Sedgwick, 2014).

Another target treatment effect is efficacy, the effect of treatment receipt on outcome (Hernán and Hernández-Díaz, 2012). This is sometimes known as the *de jure* effect (Mallinckrodt et al., 2017). Patients and clinicians wishing to make informed decisions about the management of medical conditions would often prefer to know the efficacy and safety of an intervention when taken under optimal conditions (as this will demonstrate the greatest possible effect; Revicki and Frank, 1999). Trials can be designed to answer this question and these are known as "explanatory" RCTs (Sedgwick, 2014). Such trials occur in carefully controlled circumstances and use exclusion restrictions to create a narrowed population to answer the question, for example excluding those with other

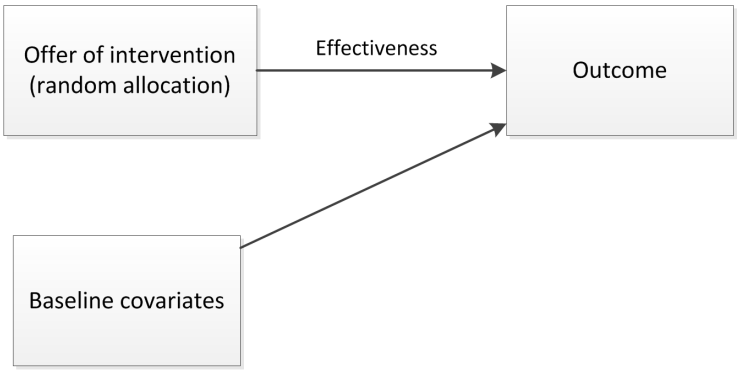


Figure 1.1: Illustration of typical randomised controlled trial demonstrating the target effect of effectiveness (the effect of treatment offer on outcome).

health conditions. These trials are said to have high internal validity, meaning that the estimated treatment effect can be attributed to the intervention and not to other factors. The cost of this is often reduced generalisability due to the strictly defined nature of the study population.

A trial where the target effect of interest is effectiveness is shown in Figure 1.1. The illustration demonstrates that this effect represents the effect of random treatment offer on outcome; covariates measured at baseline may predict outcome but they do not predict treatment offer due to randomisation. A typical trial with non-adherence is shown in Figure 1.2. This figure demonstrates what is meant by the target effect of efficacy. It shows that the offer of intervention affects what treatment participants receive. The treatment receipt variable encompasses both contamination (treatment receipt under offer of control) and non-compliance (non-receipt under offer of treatment). Treatment receipt will sometimes be referred to in this section as the exposure variable as its effect is the one of interest. The figure also shows something that is highly likely to occur: the presence of unmeasured confounding (hence the use of a circle for a latent variable) affecting the relationship between treatment receipt and outcome. All variables to the right hand side of random allocation are measured post treatment. Baseline covariates cannot be related to random allocation but they can be prognostic of outcome and treatment receipt.

In order to estimate efficacy, a number of conclusions flow from Figure 1.2. For instance, there must be an individual-level measure of treatment receipt for all participants (or at the very least a large sample) in the trial. This variable may be measured on a categorical or continuous scale and it should be used to estimate the effect of treatment receipt on

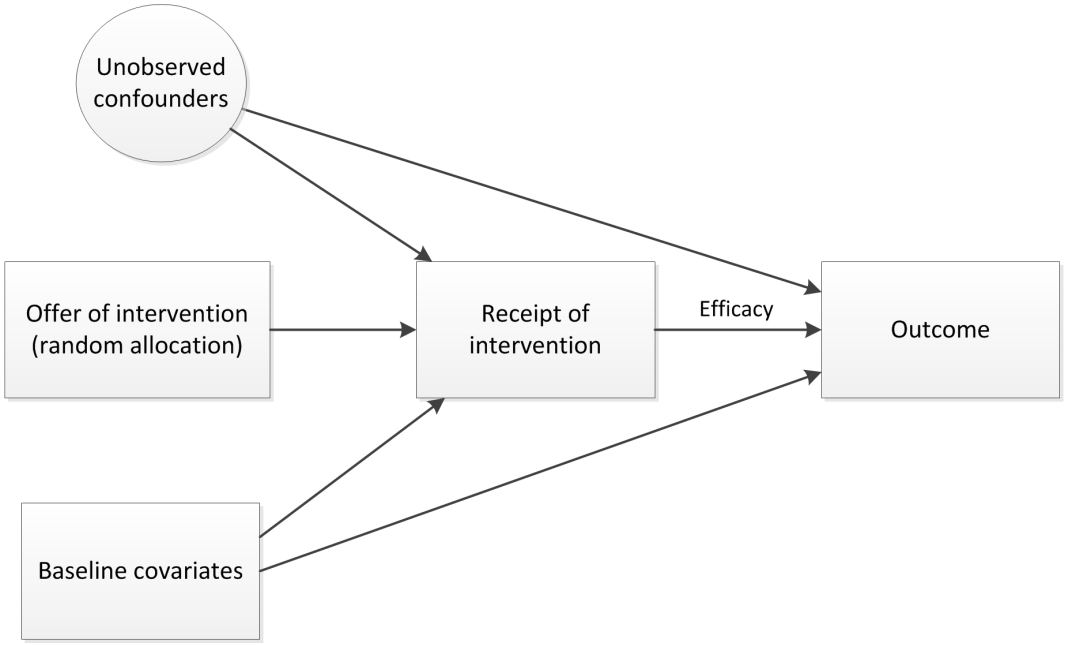


Figure 1.2: Illustration of typical randomised controlled trial with non-adherence and demonstrating the target effect of efficacy (the effect of treatment receipt on outcome).

outcome. In addition, such an estimator should be capable of adjusting for unobserved confounding between exposure and outcome, and should include prognostic baseline variables to aid precision.

Non-randomisation-based efficacy estimators

Two approaches that have been used traditionally by clinicians to estimate efficacy are the *as treated* and *per protocol* / *on-treatment* analyses. The first of these evaluates participants according to which treatment they received, irrespective of random treatment allocation. This estimator may be subject to confounding if participants' treatment receipt is associated with baseline variables that are prognostic of outcome (Hernán and Hernández-Díaz, 2012). The expectation is that these variables cannot all be measured and adjusted for in the analysis. The per protocol or on-treatment analysis estimates the effect of random treatment allocation on outcome, but only for those who adhere to allocation. Similarly to the as-treated estimator, this raises the prospect of confounding biasing the estimate if prognostic variables are predictive of treatment receipt. It has been argued that the consideration of the per protocol analysis as a supplement to the ITT approach has considerable shortcomings (White, 2005). This is primarily because the per protocol and ITT approaches are estimating distinct target effects (efficacy and effectiveness, respectively). In addition, a per protocol analysis is unlikely to be unbiased.

It is argued that in an open trial of an intervention there is no good reason to expect those who comply with protocol in the treatment and control groups to be comparable (White, 2005).

In summary, the problem with both the as treated and per protocol analyses is that they throw away the benefits of random treatment allocation. In effect these estimators treat an RCT as if it were an observational study, ignoring the bias driving treatment uptake (selection bias). These estimators risk introducing confounding unless very strong assumptions are made. Greater detail will be given about these estimators in Chapter 4.

Randomisation-based efficacy estimators

Another class of estimators allows the estimation of unbiased estimates of efficacy for those who would comply with protocol. Contrary to the as-treated and per protocol analyses these estimators are randomisation-based. A large volume of work in the last twenty years has centred around one target treatment effect in particular: the complier average causal effect (CACE). Full details of the estimation of this parameter and necessary assumptions will be given in Chapter 4. What follows here is an introduction to the methods and a summary of their application in the context of contamination in the trials literature.

CACE is the effect of treatment among those participants who would receive treatment when offered it and would not receive it when offered control. It is also sometimes known as a local average treatment effect (LATE) as this parameter refers to an effect within a subpopulation (Angrist et al., 1996), as distinct to the average treatment effect (ATE) which is the effect of treatment receipt across the whole population. The subpopulation is latent because it is not possible to determine whether a given participant belongs to this class. This group is often given the label of "compliers" (e.g. Frangakis and Rubin, 2002), although to avoid confusion with other uses of this word I will refer to these participants as "latent compliers". In order to estimate CACE it is assumed that random treatment allocation affects treatment receipt, has no direct effect on outcome other than through treatment receipt, and that there is no variable that affects both treatment allocation and outcome (this is secured by randomisation). It is also assumed that the population contains no participants whose treatment receipt is always the opposite to their allocation, i.e. no people who would not receive treatment when offered it and would receive it when offered control. This efficacy parameter can be estimated using

instrumental variable (IV) models, structural mean models, or principal stratification techniques. Full details of these estimation methods will be given in Chapter 4.

One common IV estimation approach is the two-stage least squares method. This approach regresses the clinical outcome on residuals from the regression of treatment receipt on random allocation (Angrist and Imbens, 1995). In the context of this research, it can estimate the effect of treatment amongst non-contaminators. The method is commonly applied to estimating the effect of treatment amongst compliers. For example, Dunn et al. (2003) estimated the effect of treatment amongst those who would attend a full course of psychotherapy when offered it compared to no intervention when offered control. The ITT analysis of the trial estimated effectiveness to be just under two points in the direction of benefit on the Beck Depression Inventory (BDI). The authors found that the proportion of compliance was just over half. The two-stage least squares estimation of CACE found an efficacy estimate of about 3 and a half BDI points in favour of treatment.

The main weakness of IV methods is that the assumption of no direct path from random treatment assignment to outcome is untestable. In unblinded trials, this may be a strong assumption if it is possible that the offer of treatment could affect outcome via some route other than treatment receipt, for example by improving healthy behaviour. The assumption may also be questionable when treatment receipt is dichotomised using a threshold as it is then necessary to assume that any dose of treatment below threshold has zero effect.

One of a small number of RCTs that has attempted to account for contamination in the analysis was a trial of screening in prostate cancer (Roobol et al., 2009). The authors reported that an earlier analysis of effectiveness of screening on mortality provided an estimated risk ratio of 0.80. The trial had recorded receipt of post-randomisation prostate-specific antigen tests for participants in one centre and this enabled an analysis of efficacy. The study extrapolated from this centre to all participants in the trial and estimated that 24% in the screened arm did not receive the intervention and 15% in the control arm did receive screening. The authors used profile likelihood estimation to obtain a point estimate and standard errors for efficacy, using methodology developed previously in the context of contamination (Cuzick et al., 1997). Using the estimates of non-compliance and contamination, the estimated risk ratio (of efficacy) was roughly 0.70.

The IV methodology has been extended to a continuous exposure variable, i.e. the relationship between a continuous measure of treatment receipt and outcome (Maracy and Dunn, 2011). A model must be assumed for the relationship between dose and response, for example a linear relationship between dose and treatment effect. The two-stage least squares estimator can then be used to estimate the gradient (and standard error) of this slope. It is possible to evaluate the shape of the relationship between dose and response using a sliding window approach (Burgess et al., 2014). However, this requires a sample size that is considerably larger than that found in typical RCTs in mental health.

Principal stratification is a framework that parallels IV. It makes a similar set of assumptions as described in the context of IV and is more explicit about subpopulations. It does this by dividing the target population into latent "principal strata" which are defined by what treatments participants would receive under offer of experimental intervention and under offer of control (Jin and Rubin, 2008). The drawback of both the IV and principal stratification approaches is that they reduce the effective sample size, leading to wider confidence intervals (Dunn, 2013). Thus, the cost of bias correction (defining the ITT approach as a biased estimate of efficacy) is the inflation of variance.

Comparing the efficiency of analytical methods

The report on contamination in trials of educational interventions (Keogh-Brown et al., 2007) carried out two data simulations: one that compared bias in cluster and individual randomised trials (described in Section 1.5.1) and another that compared statistical power between CRCTs with an ITT approach and individual randomised trials that estimated CACE. This section will describe the methods and results of the second round of simulations, and discuss how this type of research could be developed.

The authors compared two design approaches for addressing treatment contamination in trials using Monte Carlo simulations. In one approach they simulated CRCTs with two magnitudes of effect size, various levels of cluster size, and a fixed ICC. Cluster randomisation was assumed to prevent contamination entirely. In the other approach they simulated individual randomised trials with treatment effect sizes (same sizes as in the CRCTs) and varying amounts of contamination. They then investigated the sample sizes that would be needed in order to achieve a particular level of statistical power. They used the ITT estimator for a CRCT (with no contamination) and the CACE estimator for

an individual randomised trial (with contamination) as estimators of efficacy.

The results showed that at every cluster size that was investigated (sizes of 10, 30, 50, 100) the sample size needed for the CRCT design was greater than that of the individual randomised trial design, up to a point that was just under 30% contamination. At a cluster size of 10, 21% more participants were needed in the former trial design compared to an individual randomised trial with 20% contamination that estimated CACE. The ICC was set to 0.04 (in the CRCT trial design) for all comparisons that they made. This difference in sample size increased with cluster size, e.g. at size 50, 170% more participants were needed relative to an individual randomised trial with 20% contamination that estimated CACE. The authors commented that the individual randomised trial (CACE) approach maintained a sample size advantage over the cluster randomised design, unless expected contamination exceeded 30%. However, it is not clear what this claim was based on considering that they did not simulate more than this level of contamination. The authors also noted the similarity of these results to those of Torgerson (2001) and Slymen and Hovell (1997), which were described in Section 1.5.1. However, these studies and their simulations made different comparisons. Torgerson (2001) and Slymen and Hovell (1997) contrasted the degree of sample size inflation in CRCTs due to clustering with that in individual randomised trials due to the recovery of power as a consequence of the acceptance of contamination and dilution of the treatment effect. Specifically, both designs were interested in the effectiveness estimand (and used the ITT estimator) – the comparison was between whether the effects of clustering or contamination would lead to greater sample size inflation in order to maintain the original level of power. The data simulations of Keogh-Brown et al. (2007) instead compared methods for estimating efficacy, a fundamentally different method of addressing contamination. The CACE approach involves measuring it and estimating efficacy, as opposed to estimating effectiveness and accepting that it will dilute the treatment contrast.

A useful development of these data simulations would be to fix sample size and compare the efficiency of efficacy estimators. The simulations could be extended to a range of sample sizes (to assess estimator properties) and a range of ICCs. This would provide information on the levels of ICC, cluster size and contamination that would tip the balance between the efficiency of cluster randomisation with use of the ITT estimator and individual randomisation with use of a CACE estimator.

1.6 D6 trial

This thesis was motivated by a trial where it was suspected that some treatment contamination occurred. Diabetes-6 (D6) was a trial of nurse-delivered psychological treatment in the primary care setting for people with poorly controlled type 2 diabetes (T2D). The trial provided an opportunity to collect individual treatment receipt data and then apply randomisation-based estimators of efficacy. What follows is a background to the trial, a summary of trial's methods and then an explanation of why contamination was suspected.

1.6.1 Background

Diabetes is a disease of the metabolic system that is characterised by high blood glucose levels, which is due to failure of the transport of glucose across the cell membrane. The most common forms are type 1 diabetes, where the pancreas fails to produce enough insulin from the outset of the disease, and T2D, where cells become insensitive and eventually resistant to the effects of insulin. The risk factors for T2D are both genetic and environmental. A large increase in the prevalence of T2D has been observed in the last 20 years (Shaya et al., 2010). It has been estimated that the prevalence of T2D in the UK in 2013 was 4.5% of the population, or roughly 2,800,000 people (Holden et al., 2017). This study estimated a threefold increase in its prevalence between 1991 and 2013. This has been linked to changing lifestyle factors, in particular increasing levels of obesity (Hillier and Pedula, 2001).

Sub-optimal glycaemic control in diabetes is common. Amongst those with T2D in the USA and UK its prevalence is estimated to be around half when the threshold is defined as 53.0 mmol/mol (Shaya et al., 2010; Holden et al., 2017). This is despite the existence of evidence-based medical and educational interventions and national guidelines (Gæde et al., 2008; Davies et al., 2008; National Institute for Health and Clinical Excellence, 2017). The reasons for poor glycaemic control are multi-factorial and include psychological barriers. For example, poor self-care is associated with low psychological well-being and distress (Peyrot et al., 2005). In particular, patients with symptoms of depression are likely to have higher non-adherence to medication and poorer diet (Ciechanowski et al., 2000).

A main aim of a health service such as the NHS is to provide patients with the ability to self-

manage long-term health problems. As such, the national diabetes guidelines emphasise the need to provide care that motivates patients to improve lifestyle and therefore better control the disease (National Institute for Health and Clinical Excellence, 2017). A group of interventions that can achieve this are psychological interventions, which are defined as talking therapies that aim to identify, challenge, and replace unhelpful health beliefs, cognitions, and emotions.

Previous research has demonstrated that psychological interventions when combined with medical therapies may be promising treatments for improving glycaemic control. Alam et al. (2009) conducted a meta-analysis of 19 trials (n=1,431 participants) of psychological interventions on glycaemic control, as measured by effect on HbA_{1c}, in patients with T2D. Interventions included supportive or counselling therapy (e.g. motivational interviewing (MI)), brief psychodynamic therapy, interpersonal therapy, and cognitive behavioural therapy (CBT). Overall, the offer of psychological treatment was found to lead to a reduction in HbA_{1c} of 5.9 mmol/mol.

Psychological interventions are normally given by highly-trained clinical psychologists, making them expensive to provide. The average cost to the NHS of a course of low-intensity psychotherapy was estimated to be £493 in 2011 (Radhakrishnan et al., 2013). The expense of psychologist-delivered psychological treatment and the impracticality of providing it to a very large number of people with T2D imply a need for another approach. Evidence suggests that allied healthcare professionals can be trained to provide basic psychological interventions and that this is associated with an improvement in glycaemic control in type 1 diabetes. For example, hospital diabetes nurses have been trained to deliver diabetes-specific psychological therapy while preserving treatment fidelity (Ismail et al., 2008), and primary care nurses have successfully been trained to use motivational techniques to improve oral medication adherence in people with T2D (Hardeman et al., 2014). An RCT of nurse-led structured care, of which MI was an important part, in routine diabetes primary care found that nurses had some basic competency but this did not develop over time (Jansink et al., 2013b). There was no evidence of an effect of the intervention in comparison to usual care on levels of HbA_{1c} (Jansink et al., 2013a) after 14 months of follow-up (effectiveness estimate was 1.4 mmol/mol). In summary, there is limited evidence regarding the delivery of high fidelity psychological treatment by primary care nurses and the effect of this on outcomes of patients with T2D.

1.6.2 Trial design

D6 was a multi-centre superiority cluster RCT with two parallel treatment arms (Ismail et al., 2018). Clusters were primary care surgeries. Its rationale was to evaluate a cost-effective and practical way of competently delivering diabetes-informed psychological treatment. It tested the effectiveness of nurse-delivered MI and CBT skills for patients with T2D in the context of primary care. Recruitment was implemented in two phases as a consequence of organisational uncertainties in the run-up to the implementation of the Health and Social Care Act 2012. Patients were recruited in 2010 and the first half of 2011.

Ethical approval was granted by the King's College Hospital Research Ethics Committee (reference 09/H0808/97) and by the respective Primary Care Trusts (reference RDLSLBex 534 and 2010/403/W). Informed consent was obtained from all individual participants included in the study. The trial was registered with ISRCTN (ISRCTN75776892) on 19 May 2010.

1.6.3 Setting and target population

The study setting was five south London boroughs (Lambeth, Southwark, Lewisham, Wandsworth, and Bexley). The study population was patients with poorly controlled T2D at large urban primary care surgeries (≥6000 patients). Surgeries were invited to participate if they had a nurse providing diabetes care. Twenty-three surgeries chose to participate.

In phase one the inclusion criteria were a diagnosis of type 2 diabetes mellitus based on WHO clinical criteria, defined as the presence of the disease for at least one year during which time it was managed by diet or oral medication with no requirement for insulin therapy; age 18-79 years; presence of diabetes for at least two years; persistent suboptimal glycaemic control, defined by $HbA_{1c} \geq 69.4$ mmol/mol on two occasions, once in previous 18 months and again when being assessed for eligibility; prescription of two oral diabetes medications (metformin and one other), and/or requirement of insulin therapy. In phase two, the inclusion criterium for glycated haemoglobin was lowered to $HbA_{1c} \geq 64$ mmol/mol.

In phase one the exclusion criteria were severe mental disorders; terminal illness and severe end-stage diabetes complications; morbid obesity (BMI > 40 kg/m^2); being

housebound; no phone or internet access; lack of understanding of English; and current receipt of psychotherapy. Patients with Patient Health Questionnaire-9 depression scores > 20 and with psychotic depression or active suicidal ideation were excluded. In phase two the threshold of morbid obesity was increased to 50 kg/m^2.

1.6.4 Active intervention

Participants were offered either the D6 intervention plus standard care or standard care alone. The D6 intervention was psychotherapy comprised of elements of MI and CBT skills and was provided by general practice nurses. They were given interactive training workshops, some training caseload, and a handbook with information about key skills over a 3-month period. They were given regular supervision including feedback on training cases, group supervision, and telephone/email support. Patients were offered four 30-minute individual sessions over two months followed by eight monthly sessions. The first six were face-to-face and the last six were given in a medium agreed by the participant and nurse.

1.6.5 Comparator treatment

The comparator treatment was standard care, as recommended by NICE, and was adapted for the local population (National Institute for Health and Clinical Excellence, 2002). This was an attention control group where the intention was for patients to see general practice nurses for the same duration and number of times as those in the active intervention arm.

1.6.6 Treatment allocation

Random treatment allocation was at the level of the primary care surgery, which was done partly to avoid treatment contamination that was anticipated if a nurse were asked to provide both control and active treatments. Other reasons for using cluster randomisation included variations in local demographics and healthcare delivery, and preventing a feeling of unfairness within surgeries (which might have led to patients seeking out the treatment). Treatment assignment was in the ratio 1:1 and was done in two phases. Treatment allocation was revealed to the trial manager after practice details were entered.

The intention was for randomisation to take place after all patients had been recruited but

this led to unacceptable delays in the training of nurses. This meant that some patients were recruited after randomisation of surgeries but were kept blind to allocation until the interventions were offered in both groups. Assessors remained blind to treatment allocation until after baseline data had been collected.

1.6.7 Outcomes

The primary outcome was HbA_{1c}, which was measured at nine, 15 and 18 months after randomisation. The primary endpoint was 18 months after randomisation. Secondary outcomes were fasting lipids, blood pressure, body mass index, and psychological state (PHQ-9 and Problem Areas in Diabetes).

If HbA_{1c} was missing at follow-up, the patient's medical records were checked in case their haemoglobin level had been recorded here for another purpose. For those patients with missing HbA_{1c} at 18 months after randomisation, the missing data point was replaced by the observation from the medical records that was nearest in time (provided this data point was recorded within ninety days of the 18-month follow-up date). A similar procedure was followed in case the 15-month HbA_{1c} measure was missing. However, any missing data at this time point were only replaced if the substitute observation had not been used to replace the 18-month observation (and provided it was within ninety days of the 15-month follow-up date). The same procedure was used for any missing 9-month HbA_{1c} data.

1.6.8 Sample size

The minimal clinically significant difference in HbA_{1c} between the D6 and standard care groups at 18 months after randomisation was 10.9 mmol/mol (standardised effect size of $d = 0.55$). The sample size calculation was 432 participants. This reflected a level of statistical power of 80%, significance level of 5%, an average cluster size of 15 patients per general practice, an ICC of 0.05, an assumed participant attrition of 20%, and a loss of two practices per arm over the course of the trial.

Three hundred and thirty-four patients were recruited in 24 clusters, of which 231 patients had at least one follow-up. The average cluster size was therefore 10 patients per cluster, smaller than the assumed size of 15 patients per cluster, providing a *post hoc* power of 77%.

1.6.9 Evaluation of clinical effectiveness

The primary, effectiveness analysis used a linear mixed model where the outcome variable was HbA_{1c} at follow-up (either 15 or 18 months after randomisation) and predictor variables were treatment allocation, categorical time, an interaction between treatment allocation and time, borough, recruitment phase, and baseline HbA_{1c} (Ismail et al., 2018). The model included random effects for nurse and participant number and an unstructured covariance structure where residual error parameters were estimated separately within the treatment allocation groups. Estimated effectiveness showed no evidence of an effect of treatment offer. Using the model reported in the primary analyses, the ITT estimate at 15 months was -0.07 mmol/mol (standard error 2.64; z=-0.03, p=.98; 95% CI -5.24, 5.10) and at 18 months (primary endpoint) was -0.79 mmol/mol (SE 2.53; z=-0.31, p=.76; 95% CI -5.75, 4.18). The primary analysis did not report the ITT estimate at nine months after randomisation. However, using the same model as applied to the other time points but this time including the nine-month data, the estimate was -0.18 mmol/mol (SE 2.46; z=-0.07, p=.94; 95% CI -5.00, 4.64).

1.6.10 Treatment contamination

It was hypothesised that there may have been some receipt of active intervention amongst participants in the standard care trial arm. There was anecdotal evidence that one of the control group nurses provided additional support to patients as a consequence of participation in the study, although it was not known to what extent this was grounded in the principles of psychological treatment. It has also been suggested that the differences between the trial arms in terms of treatment received were smaller than expected. The fidelity analysis that was performed as part of the primary assessment of the trial used the Motivational Interviewing Treatment Integrity (MITI) version 3.1.1 to evaluate the fidelity of MI (Moyers et al., 2005, 2010), and the Behavioural Change Counselling Index (BECCI) to assess the delivery of CBT skills (Lane et al., 2005). Further information about these scales will be given in Chapter 3. Treatment sessions of 33 patients in the D6 intervention group and 36 in the standard care group (3.4% of all sessions) were rated (Ismail et al., 2018).

The results of the fidelity assessment are shown in Table 1.1. There was evidence for a higher proportion of open questions in the D6 intervention group and a larger reflection/question ratio in the standard care group. There was no evidence of a difference

between the trial arms for the other MITI domains and BECCI Practitioner Score. The fidelity assessment showed that there may have been some contamination of the control group. For example, the MITI Global Spirit and Global Empathy scores showed that nurses may have been using psychological treatment techniques to a limited extent.

Table 1.1: Mean scores of treatment fidelity scales (MITI and BECCI) by treatment group from D6's primary assessment.

Domain	Standard care group	D6 intervention group	p-value†
MITI			
Global Spirit (mean; SD)	2.87 (0.87)	3.23 (1.13)	.14
Global Empathy (median; IQR)	2.50 (2.00-3.00)	3.00 (2.00-4.00)	.19
Percent Complex Reflections (mean; SD)	40.4 (17.4)	35.2 (19.8)	.25
Percent Open Questions (mean; SD)	24.9 (10.0)	36.4 (17.3)	<.01
Reflection/question ratio (median, IQR)	0.74 (0.53-1.19)	0.57 (0.42-0.72)	.03
Percent MI-Adherent (mean; SD)	53.6 (27.6)	58.4 (31.8)	.51
BECCI			
BECCI Practitioner Score (mean; SD)	1.12 (0.55)	1.33 (0.56)	.12

† Based on result of either a t-test or Mann-Whitney U test.

The anecdotal evidence of the standard care nurse who was particularly aware of psychological treatment techniques, the results of the primary analysis treatment fidelity assessment, and the small estimate of treatment effectiveness suggested that there may have been some treatment contamination. For this reason, it was decided to measure treatment fidelity for as many participants as possible and then to account for any contamination or non-compliance in an efficacy analysis of the trial. I will present the treatment fidelity assessment in Chapter 3 and the efficacy analysis in Chapter 7.

1.7 Other trials motivating this research

I obtained data from three other mental health trials that motivated the research. These were clinical trials of complex interventions that used cluster randomisation to avoid contamination. They were REFOCUS, CONMAN and a trial of systematic assessment of

need and care planning in severe mental illness. They were used primarily to provide information on typical levels of parameters relating to cluster randomised trials (intraclass correlation coefficient, cluster size). The trial of systematic assessment of need and care planning applied randomisation at both cluster and individual participant levels. This dataset provided an opportunity to assess the impact of contamination avoidance (through use of cluster randomisation) on outcomes (Chapter 2).

Recovery intervention within community-based mental health teams (REFOCUS)

REFOCUS (ISRCTN02507940) was a cluster RCT of a pro-recovery intervention for people with psychosis and used treatment allocation at the level of clinical team (Slade et al., 2015). The intervention consisted of training and reflection activities, a manual and the practice of partnership working with the aim of encouraging clinicians to promote recovery in the way they treated patients. The primary outcome was a patient-reported questionnaire of recovery and was measured 12 months after baseline.

Contingency management for people with heroin dependence and needing hepatitis B vaccination (CONMAN)

CONMAN (ISRCTN72794493) was a cluster RCT of contingency management for people with heroin dependence and needing hepatitis B vaccination and applied randomisation at the level of drug treatment clinic (Weaver et al., 2014). The primary outcome was completion of three vaccinations within 28 days and was therefore binary.

Systematic assessment of need and care planning in severe mental illness

The systematic assessment of need and care planning trial (no ISRCTN reference) tested a complex care intervention for people with schizophrenia, bipolar disorder, depression, or delusional disorder (Marshall et al., 2004). The trial comprised two subtrials in which treatment was allocated either to individual participants or to clusters of patients that were defined by who their care coordinator was.

1.8 Aims of the thesis

The primary research objective was to compare the efficiency of two competing trial design options that address the problem of contamination and estimate efficacy. The comparison built on the work by Keogh-Brown et al. (2007) by comparing similar design

methods. The first option was cluster randomisation at the level at which contamination was expected (i.e. preventing any contamination) with an estimator of the effect of treatment receipt on outcome that accounted for clustered data. The second option was individual randomisation, acceptance that contamination would take place, measurement of treatment receipt for all participants, and use of a randomisation-based efficacy estimator of the effect of treatment on outcome within the subpopulation of compliers. I aimed to make the findings from the comparison of these design options accessible to those planning trials in the form of an online decision support tool.

The secondary research objective was to summarise and develop randomisation-based estimators of efficacy in a randomised trial with contamination, where treatment receipt could be measured on a binary or continuous scale. Having collected individual treatment receipt data for participants in D6 as part of this research project, the second part of this objective was to apply these estimators to evaluate the efficacy of the D6 intervention.

The tertiary research objective was to review the problems and solutions associated with contamination in mental health trials of complex interventions. The aims were to assess the processes driving contamination, the typical extent of the problem, what researchers do in order to mitigate it, and the quantity of contamination bias within trials that used both individual- and cluster-level randomisation.

1.9 Thesis overview

In this thesis I will describe methods that can be used to address the problem of contamination in the design and analysis of trials of complex interventions in mental health. The following chapters will start with a large scoping review of contamination in trials (Chapter 2). This will explore the drivers of contamination and the solutions trialists use to address the problem. It will also describe the extent of contamination and evaluate whether there is evidence that its occurrence is related to estimated treatment effect sizes. Chapter 3 will assess treatment fidelity in the D6 trial, the study that motivates much of this research due to the anticipated presence of contamination. The fidelity assessment will use a sample of data collected for this doctorate by two clinical psychologists who were trained in the BECCI and MITI scales. In Chapter 4 I will summarise existing analytical methods for estimating efficacy in trials with contamination. I will also develop a novel estimator of efficacy when non-compliance in the active arm and contamination

in the control arm are measures on a continuous scale. In Chapter 5 I will compare the efficiency of two competing trial design options for evaluating efficacy in the presence of contamination using Monte Carlo simulations:

A. Treatment allocation by cluster randomisation with clusters defined at the level at which contamination takes places together with an estimator of the effect of treatment receipt on outcome that accounts for clustered data.

B. Allocation at the participant level, measurement of treatment receipt in all participants, and estimation of efficacy using a randomisation-based estimator to target CACE.

These competing design/analysis options will be investigated for both binary and continuous measures of treatment receipt. In Chapter 6 I will describe the development of an online decision support tool to help those planning trials choose between the two design options for dealing with contamination. This will use the results from the Monte Carlo simulations. In Chapter 7 I will apply the existing estimators (of CACE) and the novel one that is able to estimate efficacy with a continuous measure of treatment receipt in both trial arms to the D6 trial. Finally, in Chapter 8 I will summarise the findings and describe the novelty of the research. I will evaluate the limitations of the methods and explore possible areas for further enquiry.

This methodological research is motivated by D6 and the three other trials in mental health. Data from all four of these trials were used to guide parameter choices when comparing two trial design options for addressing contamination (Chapter 5). A list of these trials and the uses of their datasets in this research can be found in Table 1.2.

Table 1.2: Sources of data in the PhD and uses of the datasets in this research.

Source of data	**Description**	**Use(s) of datasets**
D6 trial - primary outcome (HbA_{1c}) dataset	Data used in the primary analysis of trial of psychological treatment for people with poorly controlled type 2 diabetes.	1. Selection of parameter levels in data simulations comparing two trial design options for addressing contamination (Chapter 5). 2. Efficacy analysis of D6 trial (Chapter 7).

D6 trial - fidelity dataset	Treatment receipt data using two scales that measure the fidelity of psychological treatment. These data were collected as part of the PhD (coding of audiotapes by clinical psychologists).	1. Fidelity assessment of D6 trial (Chapter 3). 2. Construction of treatment receipt variable for efficacy analysis of D6 trial (Chapter 7).
REFOCUS trial	Cluster RCT of recovery promotion intervention (training, reflection activities, manual, partnership working) for people with psychosis. Clusters were clinical teams.	Selection of parameter levels in data simulations comparing two trial design options for addressing contamination (Chapter 5).
CONMAN trial	Cluster RCT of contingency management for people with heroin dependence and needing hepatitis B vaccination. Clusters were drug treatment clinics.	Selection of parameter levels in data simulations comparing two trial design options for addressing contamination (Chapter 5).
Trial of systematic assessment of need in severe mental illness	RCT of complex care intervention for people with mental health disorders. Treatment allocation was at cluster and individual levels for different parts of the trial.	1. Selection of parameter levels in data simulations comparing two trial design options for addressing contamination (Chapter 5). 2. Comparison of treatment effects between cluster and individual randomised subtrials in the scoping review (Chapter 2).
Scoping review of contamination	Results of scoping review of problems and solutions associated with contamination in trials of complex interventions. This was conducted as part of the PhD (Chapter 2).	Selection of parameter levels in data simulations comparing two trial design options for addressing contamination (Chapter 5).

Chapter 2

Scoping review of problems and solutions associated with contamination

2.1 Background and aims

The processes leading to contamination in mental health trials have never been reviewed comprehensively and the literature is unclear about their relative frequencies. This knowledge is necessary in order to plan what steps should be taken to address the problem. In mental health trials, some of the methods that researchers use to minimise or prevent contamination are either little known or poorly formalised within the literature. Knowledge of these points could provide those planning trials with additional tools for addressing small amounts of contamination without resorting to the use of cluster randomisation and the associated increase in sample size requirement.

This chapter addresses the tertiary objective of the research, which was to review the problems and solutions associated with contamination in mental health trials of complex interventions. The aims of this chapter were to identify the processes that are considered to lead to contamination in trials of complex interventions in mental health, to quantify typical levels of contamination, to summarise what researchers do in order to prevent or mitigate it, and to compare treatment effect estimates within trials of complex interventions that used both cluster- and individual-level treatment allocation to quantify the contamination bias. In addition, the chapter aimed to summarise relevant parameters to be used in later work that will compare different design and analysis options by

simulation using realistic parameter choices. Here I describe a comprehensive scoping review that addresses these points.

As mentioned before in Section 1.5, I categorise possible solutions to address contamination into:

i. Statistical design methods – this includes structural design methods such as cluster randomisation, sample size inflation, and participant preference designs,

ii. Trial conduct solutions – this relates to methods that can be used in the running of the trial to reduce exposure to active treatment in the control arm,

iii. Analytical approaches – this consists of using a measure of treatment receipt to estimate efficacy.

The review I conducted included a small number of trials that allocated treatment at both cluster and individual levels (including the trial of systematic assessment of need and care planning in severe mental health; Marshall et al., 2004). This provided an opportunity to assess the extent to which contamination impacts on outcomes.

This chapter will provide a description of the review's methodology, summaries of assessments of evidence of trial bias, details of contamination processes and their frequencies, solutions used to prevent contamination in mental health research, an investigation of the evidence of the effect of contamination on estimates of treatment effect from four trials that used random allocation at both the levels of the individual and the cluster, and summaries of relevant parameters.

The work presented here has been submitted for publication as an article to BMC Medical Research Methodology.

2.2 Methods

2.2.1 Type of review

I carried out a scoping review of trial design and conduct methods in RCTs of complex interventions in mental health. This type of review was chosen on the basis that my objectives were to summarise researchers' perceptions of and solutions to a trial design problem where there is limited literature and potentially highly heterogeneous evidence.

2.2.2 Eligibility criteria

All articles published between 2000 and 2015 that describe contamination in mental health trials of complex interventions were screened using full texts and were assessed using five inclusion criteria. First, the text described a trial purporting to have used random allocation. Second, the intervention was complex, which in this review meant it comprised multiple components. It was not possible to assess whether these elements acted together to provide some added benefit (as per MRC guidance definition) so I used this general and therefore wide definition for this. Third, the publication gave some information about the process leading to, amount of, or solution used to counter treatment contamination. Fourth, the abstract and main body of the article were written in English. And finally, the trial was related to mental health, psychology, or psychiatry – this meant that a minimum of one of the target population, intervention, or primary outcome was directly related to one of these fields. Many trials in these fields test unblinded treatments where the suspicion is that they may be subject to contamination. The scoping review was limited to these areas of medicine for this reason and because of the apparent gap in the literature surrounding contamination in these fields.

2.2.3 Information sources

The search for contamination in RCTs of complex interventions in mental health was done using the Ovid platform and included the databases Medline, Embase and PsycInfo. Articles that were published between January 2000 and April 2015 were searched. Results were restricted to those articles published after 2000 because this was the year when the MRC framework paper on complex interventions was first published (Campbell et al., 2000). The publication of this framework marked the point at which the design and evaluation of complex interventions were formalised.

2.2.4 Search

Randomised controlled trials were searched for using the sensitivity-maximising 11-step process recommended by the Cochrane Collaboration (Lefebvre et al., 2011). The search terms “contamination” and “spillover” were included in the procedure. it was found that searching for these words without qualifiers produced a great number of articles that were not relevant to the review. Instead, these terms were combined with the words “treatment”, “arm”, “control”, “group”, “outcome”, “trial”, “patient”, or “intervention”

with a maximum gap of six other words between the two. Synonymous terms for complex interventions that were used included all combinations of “multicomponent”, “multifaceted”, “psychosocial”, and “behavioural”, with “interventions”, “treatments”, and “training”. In order to increase sensitivity, the adjective and the noun could be separated by three other words. Also included were the terms “psychotherapy” and “therapist”. All terms were searched for in the main body of the text.

The search was restricted to articles that mentioned “mental health”, “psychology”, or “psychiatry”. As an additional method of improving specificity, certain terms were excluded, e.g. “blood contamination”, “microbial contamination”, “device”, “vaccine”, “microbial”, ‘antimicrobial”, “genes“, “genetic”, “screening”, “decision aid”, and “decision support”. Many of these terms were only used as exclusion criteria if they were found within the title, subject heading, or abstract. Full details of the search procedure can be found in Appendix A.1.

2.2.5 Study selection

Duplicates were removed from the set and the remaining articles were assessed for each of the exclusion criteria. Any potentially relevant article that was referred to by a paper in the results of the search and was not already in the set was followed up. If the article was judged to have met the inclusion criteria it was included in the set and the full text was reviewed. In order to assess the reliability of study selection, a second reviewer (another PhD student, Ruth Knight) re-screened 70 articles (11%).

2.2.6 Data collection process

Any studies that were included in the review and featured substudies that used both cluster- and individual-level treatment allocation were reviewed as two separate subtrials because of the different contamination processes and methods used to address these. Treatment effect sizes were extracted for these subtrials that reported effects separately depending on the level of treatment allocation. Data from any such studies that did not report results at the different levels of treatment allocation were obtained from the authors in order to allow the comparison.

2.2.7 Data items

Abstracted data included an assessment of bias, summaries of trial design (e.g. study population, intervention, primary outcome, unit of treatment allocation), details about contamination (e.g. how it was thought to take place, its quantity, steps taken to avoid it), and records of trial summaries (e.g. extent of clustering, power, sample size, treatment effect). A full list of data extracted for each trial can be found in Appendix A.2. In order to assess the reliability of data abstraction, the second reviewer (RK) re-extracted data from 20 articles (8%) using the same procedure described above.

2.2.8 Risk of bias in individual studies

The review of trial bias included recording the "Jadad score" (a single item measure of methodological quality of RCTs; Jadad et al., 1996) and all the domains of the Cochrane Collaboration's classification scheme for bias (Higgins et al., 2011). In addition to these, some other domains that were pertinent to cluster randomised trials were used. These included whether randomisation occurred after participant consent was obtained, baseline measures were completed before randomisation, baseline outcome measurements were similar across trial arms, other clinical and demographic characteristics were similar across arms, and whether attrition was similar in the arms. These additional assessments of bias were based on outcomes used in reviews of CRCTs (Puffer et al., 2003). Fuller descriptions of items relating to risk of bias can be found in Appendix A.2.

2.3 Results

2.3.1 Reliability

At the screening stage agreement was 71%; all discrepancies were discussed and subsequently resolved. Agreement was 81% for all assessments of bias, and 82% for details of contamination processes.

2.3.2 Summary of trials

Two hundred and thirty-eight trials were identified as meeting the eligibility criteria. This included seven trials that were referred to by an article in the main search and were found to meet the eligibility criteria. The results of the implementation strategy and numbers of exclusions are summarised in Figure 2.1. Details of the 238 articles in

the review are given in Table 2.1. The table shows that the overwhelming majority of articles described the primary analysis of an RCT, with the trials based in either North America or Western Europe, and were late phase (i.e. not pilot or feasibility trials). Most target populations were adult patients and the most commonly targeted conditions were depression, substance abuse, and psychosis. The two most common interventions were cognitive behavioural therapy and care management; there were many small categories. A full list of references can be found in Appendix A.3.

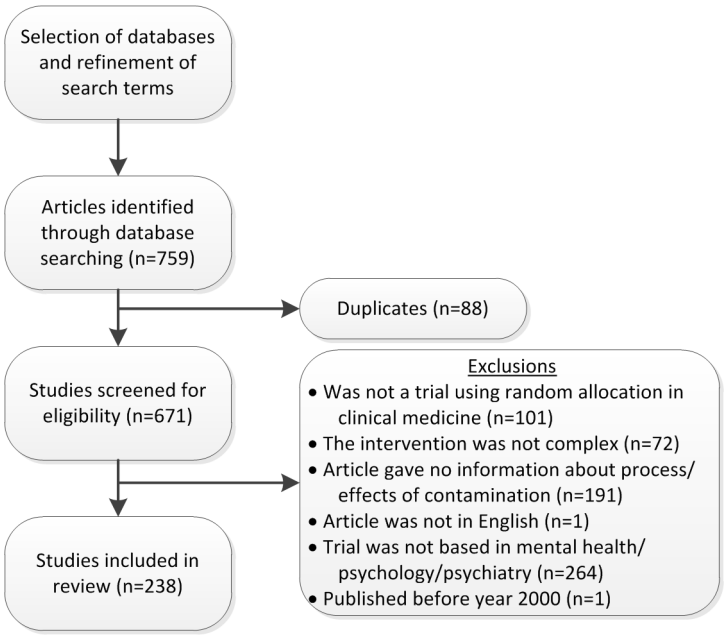


Figure 2.1: Flow diagram for searching for relevant articles (articles could be excluded for more than one reason).

2.3.3 Summary of assessment of bias

Summaries of assessments of bias using the Jadad scale, items recommended by the Cochrane Collaboration, and items aimed at identifying possible bias in cluster randomised trials are reported in Table 2.2. The table demonstrates the potential for bias split by the level of treatment allocation. In general, it shows that the greatest potential for bias arose due to incomplete outcome data being inadequately addressed, differences in attrition between trial arms, and randomisation occurring before consent was obtained and before baseline measures were completed. More potential for bias was found in cluster randomised trials when assessing whether randomisation took place after consent and after baseline measures were completed, whether outcome assessment was blind, and whether attrition was similar between trial arms. There was some suggestion that

Table 2.1: Summary of characteristics of trials (n=238 trials).

Variable	Level	Number of articles
Type of article (n)	Results of primary analysis of clinical trial	232 (97.5%)
	Design / protocol of clinical trial	3 (1.3%)
	Results of secondary analysis of clinical trial	3 (1.3%)
Year (n)	2000-2004	48 (20.2%)
	2005-2009	87 (36.6%)
	2010-2015	103 (43.3%)
Country of origin (n)	USA	102 (42.9%)
	UK	42 (17.6%)
	Netherlands	19 (8.0%)
	Canada	14 (5.9%)
	Australia	11 (4.6%)
	Other	50 (21.0%)
Target population (n)	Adult patients	175 (73.5%)
	Children / adolescent patients	45 (18.9%)
	People at risk	6 (2.5%)
	Workers	12 (5.0%)
Target condition (n)	Depression	31 (29.2%)
	Substance abuse	18 (17.0%)
	Psychosis	14 (13.2%)
	Neurodegeneration	13 (12.3%)
	Anxiety	6 (5.7%)
	ADHD	5 (4.7%)
	Others	19 (17.9%)
	No single target condition	132 (55.5%)
Intervention (n)	Cognitive behavioural therapy / CBT skills	33 (13.9%)
	Care management / interdisciplinary care	27 (11.3%)
	Education	21 (8.8%)
	Motivational interviewing / motivational enhancement therapy	19 (8.0%)
	Other psychotherapy / counselling	16 (6.7%)
	Assessment and feedback	8 (3.4%)
	Parenting interventions	8 (3.4%)
	Others	106 (44.5%)
Phase (n)	Early (pilot and feasibility trials)	30 (12.6%)
	Late	208 (87.4%)
Level of treatment allocation (n)	Participant level	145 (60.9%)
	Cluster level	93 (39.1%)
Participant sample size (median; range)	Participant-level allocation	143 (16-14910)
	Cluster-level allocation	251 (13-6076)
Cluster size in CRCTs (median; range)		10 (3-200)

individual-randomised trials were more prone to bias when assessing whether incomplete outcome data were adequately addressed (disregarding those trials where this was unclear).

Table 2.2: Summary of assessment of bias in trials.

Variable	Level	Trials with individual-level randomisation (n)	Trials with cluster randomisation (n)
Jadad score (possible range of 0-5; higher scores indicate lower likelihood of bias)	0	2 (1.4%)	0 (0%)
	1	30 (21.1%)	25 (27.2%)
	2	63 (44.4%)	38 (41.3%)
	3	47 (33.1%)	29 (31.5%)
Allocation sequence adequately generated	Yes	84 (57.9%)	46 (49.5%)
	No	3 (2.1%)	1 (1.1%)
	Unclear	58 (40.0%)	46 (49.5%)
Allocation sequence adequately concealed	Yes	94 (64.8%)	52 (55.9%)
	No	3 (2.1%)	2 (2.2%)
	Unclear	48 (33.1%)	39 (41.9%)
Randomisation after consent obtained	Yes	124 (85.5%)	24 (25.8%)
	No	5 (3.4%)	45 (48.4%)
	Unclear / NA	16 (11.0%)	24 (25.8%)
Randomisation after baseline measures were completed	Yes	64 (44.1%)	19 (20.4%)
	No	20 (13.8%)	49 (52.7%)
	Unclear	61 (42.1%)	25 (26.9%)
Baseline outcome measurements similar across trial arms	Yes	117 (80.7%)	73 (78.5%)
	No	6 (4.1%)	9 (9.7%)
	Unclear / NA	22 (15.2%)	11 (11.8%)
Baseline demographic characteristics similar across trial arms	Yes	125 (86.2%)	74 (79.6%)
	No	3 (2.1%)	14 (15.1%)
	Unclear / NA	17 (11.7%)	5 (5.4%)
Knowledge of allocation adequately concealed	Yes	0 (0%)	1 (1.1%)
	No	145 (100%))	92 (98.9%)
Blinded outcome assessment	Yes	73 (50.3%)	35 (37.6%)
	No	10 (6.9%)	13 (14.0%)
	Unclear / NA	62 (42.8%)	45 (48.4%)
Incomplete outcome data adequately addressed	Yes	48 (33.1%)	41 (44.1%)
	No	50 (34.5%)	32 (34.4%)
	Unclear / NA	47 (32.4%)	20 (21.5%)
Similar attrition between trial arms	Yes	98 (67.6%)	52 (55.9%)
	No	29 (20.0%)	26 (28.0%)
	Unclear / NA	18 (12.4%)	15 (16.1%)

2.3.4 Processes driving contamination

There were perceived to be five main processes that led to contamination. The first two processes, staff delivering the active intervention in the control arm (n=85, 36%) and communication between trial arms (n=80, 33%), were the most common. Staff delivering the active intervention in the control arm happened either due to a given clinician delivering both the active and control treatments (n=79, e.g. Barkhof et al., 2013) or due to control participants being exposed to the intervention as a consequence of

clinicians, who were not directly involved in providing the treatment, treating participants in both arms and thereby potentially learning about the active intervention and passing this on to participants in the control arm (n=6, e.g. Beck et al., 2002). The other main contamination process was communication between individuals in different trial arms. This could be either at the level of the clinician (n=20, e.g. Johnson et al., 2007), participant (n=58, e.g. Ersek et al., 2008), or both (n=2). Communication between providers of interventions was often a worry in environments in which the people giving the treatment worked closely together, for example GP surgeries, hospital units, and schools. Communication between participants was thought to be most likely in environments in which participants came into close contact. Examples of this included interaction between participants who were family members, patients in a waiting room, school children, employees working on the same site, and university students. Particular healthcare settings that were thought to be highly likely to foster communication were antenatal clinics/childbirth classes, specialist clinics (e.g. substance misuse, dialysis), and wards for those admitted to hospital.

There were perceived to be three other, more minor processes that drove contamination. First, participants switching clinicians (n=4, 2%, e.g. Cooper et al., 2013), where control participants were treated by multiple clinicians of whom one was trained in the active intervention. Second, participants seeking treatment outside the trial (n=6, 3%, e.g. Stuifbergen et al., 2010). And finally, what I have called background noise, where the treatment already existed to some extent within the healthcare system (n=5, 2%, e.g. Becoña and Vázquez, 2001). Fifty-nine articles did not provide information about the contamination process.

The five processes driving contamination are shown as a scheme in Figure 2.2. This figure also includes my interpretation of whether cluster randomisation could be used to prevent particular types of contamination. Cluster randomisation could be used to prevent contamination due to communication between clinicians or participants, provided that clusters are selected at a level that is high enough to prevent this contact. It may be able to prevent contamination due to crossover of staff between treatment groups. If contamination is due to the transfer of information about the treatment by clinicians who are not directly involved in providing active intervention, cluster randomisation may prevent contamination if clusters are constructed at a high enough level. Its use may be capable of preventing contamination due to participants switching clinicians (provided

clusters are constructed at a high enough level) but will not prevent contamination due to participants seeking treatment outside the trial or due to the treatment being already available within the healthcare system.

2.3.5 Quantity of contamination

Twenty-nine studies (11%) attempted to quantify contamination. Twenty-five trials measured individual-level contamination on a binary scale and summaries of these quantities are given in Table 2.3. The median level of contamination was 13% (IQR 5-33%).

Table 2.3: Quantity of treatment contamination in trials where participants could either receive or not receive treatment (i.e. treatment receipt measured on binary scale).

Reference	Control treatment	Active intervention	Measure of contamination	Contamination (n)
Aveyard et al. (2007)[i]	Basic behavioural support for smoking cessation	Behavioural support for smoking cessation	Nurse visit (1st extra); Telephone call; Nurse visit (2nd extra)	12/469 (3%) 12/469 (3%) 5/469 (1%)
Barton et al. (2004)[i]	No treatment	Mammography education (pamphlet and videotape) focusing on anxiety	Patient recall of: Pamphlet; Videotape	 9%; 1%
Bernstein et al. (2005)[c]	No treatment	Cognitive behavioural therapy	Service Questionnaire of anxiety treatment	0/24 (0%)
Borland et al. (2013)[i]	Minimal information	Behavioural support	Patients reporting use of extensive behavioural support	45/378 (12%)
Clarkson et al. (2009)[i]	Routine care	Self-efficacy education	Participants reporting use of electric toothbrush	9/113 (8%)
Clarkson et al. (2009)[c]	Routine care	Self-efficacy education	Participants reporting use of electric toothbrush	9/180 (5%)
Courneya et al. (2003)[c]	Group psychotherapy	Group psychotherapy and exercise programme	Patient-reported exercise	10/45 (22%)
Dilley et al. (2007)[i]	Usual care	Cognitive counselling	Patient-reported receipt of counselling	45/158 (29%)
Forchuk et al. (2005)[c]	Usual care	Transitional discharge from hospital	Patient-reported receipt of peer support and staff contact	27%

Heirich and Sieck (2000)[i]	Health education	Proactive follow-up counselling	Patients requesting personal counselling	56%
Johnson et al. (2007)[c]	Usual treatment	Clinical training in dual diagnosis of psychosis and substance misuse	Patients not taken on by trained case manager	19/105 (18%)
Khumalo-Sakutukwa et al. (2008)[c]	Standard HIV voluntary counselling and testing	HIV counselling, testing and self-management	Participants seeking out treatment from intervention centres	1%
Lamers et al. (2010)[i]	Usual care	Nurse-led minimal psychological intervention (MPI)	Patients who reported knowledge of MPI	9/178 (5%)
Lee and Gay (2011)[i]	Attention control	Sleep hygiene package	Patient-reported use of bassinet	33/46 (72%)
Lee and Gay (2011)[c]	Attention control	Sleep hygiene package	Patient-reported use of: Bassinet; White noise device; Low lighting	 47/75 (62%) 11/75 (14%) 27/75 (36%)
Merritt et al. (2007)[c]	No intervention	Postcards with information about depression	Patients reporting having seen the postcards	7/78 (1%)
Moadel et al. (2012)[i]	Standard care	Smoking cessation group support and encouragement	Patients reporting discussion with active intervention patients; Patients reporting familiarity with programme's strategies	6% 17%
Mohr et al. (2011)[i]	Treatment as usual	Cognitive behavioural therapy	Patients who had contact with non-study therapist	18/44 (41%)
Phillips et al. (2014)[c]	Routine public health practice	Community engagement in healthy eating	Participants reporting participation in intervention programme	1%
Saitz et al. (2013)[i]	Usual care	Chronic care management (multidisciplinary care coordination; motivational therapy; counselling)	Patients who received a session of motivational enhancement therapy	9/281 (3%)
Shemilt et al. (2004)[c]	No funding for breakfast club	Funding for school-based breakfast club	School pupils with school breakfast club	77%

Stewart-Brown et al. (2004)[i]	No intervention	Incredible Years (parenting techniques) training	Participants attending community-based parenting programme	4/44 (9%)
Waghorn et al. (2014)[i]	Enhanced routine mental health case management	Supported employment and specialist illness management	Patients opting to transfer to intervention after 6 months	28/102 (27%)
Walpole et al. (2013)[i]	Social skills training	Motivational interviewing (MI)	Patients whose treatment was MI adherent	37%
Wells et al. (2000)[c]	Usual care	Quality improvement therapy (CBT) and medications (assessment and education)	Receipt of speciality counselling within 6 months	13%

[i] Trial using individual-level randomisation
[c] Trial using cluster-level randomisation

Contamination was measured using a continuous scale in four trials. These were three trials of cognitive behavioural therapy and one of cognitive analytic therapy. One created a treatment fidelity scale and asked participants in each trial arm (behavioural weight control instructions, cognitive behavioural therapy, standard counselling) about their knowledge of all three treatments at the beginning and end of treatment (Perkins et al., 2001). The subscales showed high knowledge of behavioural weight control in the group allocated to receive behavioural weight control instructions (mean change of 1.1 compared to 0.5 and 0.5 in cognitive behavioural therapy and standard arms), high knowledge of cognitive behavioural therapy in those allocated to receive this (mean change of 1.6 compared to 0.0 and 0.8 in behavioural weight control and standard groups), and high knowledge of standard intervention in the control group (mean change of 0.5 compared to 0.1 and 0.1 in behavioural weight control and cognitive behavioural therapy arms). This seemed to indicate receipt of treatment in the control arm. Three RCTs showed negligible evidence of treatment contamination. Of these RCTs, one used a cognitive behavioural therapy adherence scale (adapted CTACS) to record compliance and contamination in the active intervention and control arms (Thorn et al., 2011). The CTACS means were 98.0 and 98.8 in the cognitive behavioural therapy and education intervention (control) arms respectively, indicating that contamination did not occur. Another trial found that the family-focused cognitive behavioural therapy (FCBT; active intervention) group scored higher than the traditional child-focused cognitive behavioural therapy (control) group on two scales, Family Focus (mean = 4.90 and 1.55) and Parenting Style Focus (mean = 4.75 and 1.00; Wood et al., 2006). This suggested

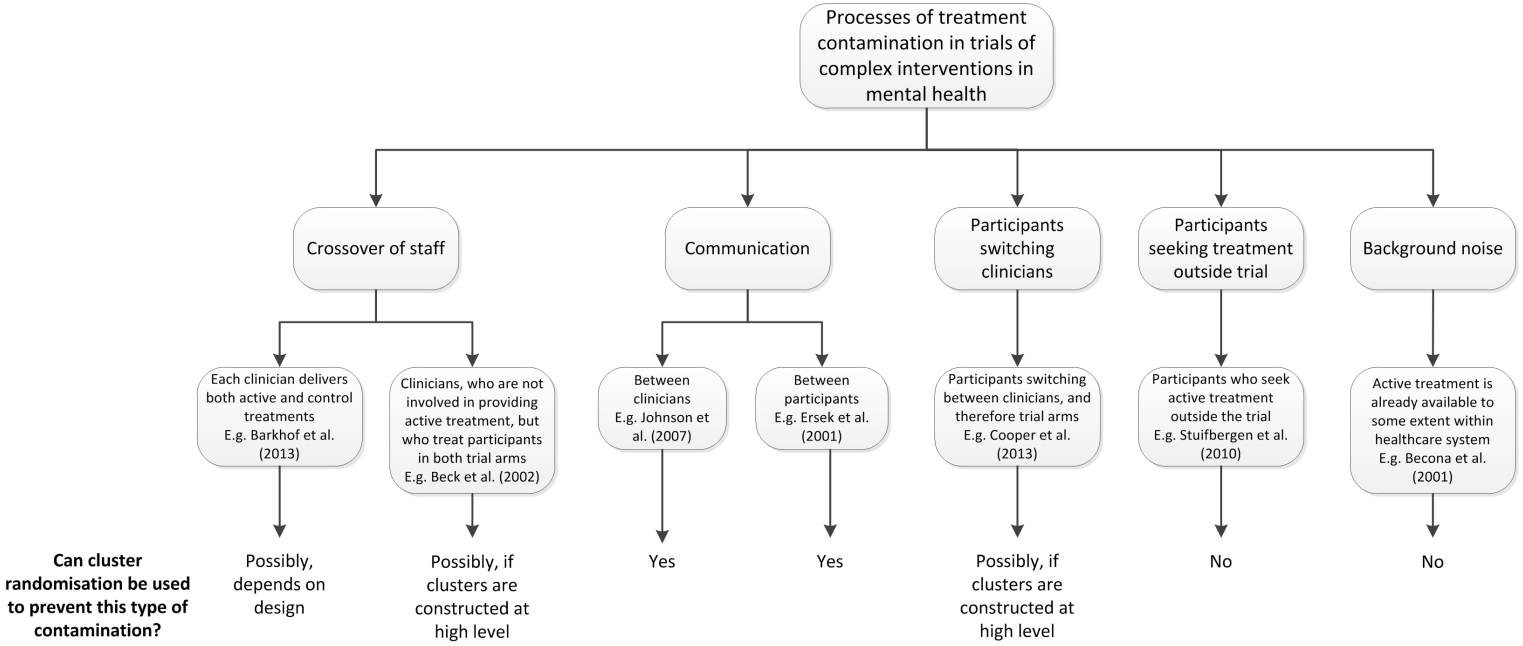


Figure 2.2: Processes driving treatment contamination in trials.

that only the FCBT group incorporated family and parenting interventions and therefore that there was little evidence of contamination. The fourth trial used a scale to measure the fidelity of the control intervention, which was good clinical care (Chanen et al., 2008). This scale included a subscale for cognitive analytic therapy and the mean for this was very low: 0.52 (SD 0.11). This represented negligible contamination.

2.3.6 Solutions used to counter contamination

Methods that were used to counter contamination are summarised in three categories: statistical design, trial conduct, and analysis methods. Statistical design includes the use of cluster randomisation, where clusters are chosen based on groups of participants who are thought potentially to become contaminated by direct or indirect links (e.g. via a shared therapist). One trial inflated the sample size in order to account for reduced statistical power caused in part by contamination bias (Dobscha et al., 2009). The great majority of other methods for preventing contamination were aspects of trial conduct, such as recruitment of more clinicians to ensure that each clinician only delivered one of the interventions. In terms of analysis methods, one trial used per protocol analysis, meaning that participants whose treatment was contaminated were dropped from the analysis (Pfiffner et al., 2007). This review found no trials that addressed the problem of contamination by using methods from the causal inference field.

A summary of trial conduct solutions that were used to avoid treatment contamination can be found in Table 2.4. These solutions have been presented in an order that matches the processes of contamination described in the earlier section on this. The majority of solutions used to prevent contamination related either to preventing staff delivering the active intervention in the control arm or preventing communication between clinicians or participants.

Investigators were concerned about contamination during data collection in four trials and aimed to prevent this by minimising interaction between researchers and participants (Chan et al., 2013; Chochinov et al., 2011; McLaughlin et al., 2005; Tiwari et al., 2005). Another temporally separated the control and active treatments with data collection following each. This meant that treatment could only influence data from active intervention participants (Alessi et al., 2005).

Table 2.4: Trial conduct solutions to treatment contamination. Ninety-two trials (out of 238) described a trial conduct solution.

Process driving contamination	Trial conduct solution	Number of trials
Clinicians deliver both active and control treatments	Recruiting groups of clinicians, each one of which is responsible for a single treatment	16
	Monitoring contamination using supervision/therapy session recordings	10
	Formalising differences between interventions, e.g. using structured manual during therapist training	6
	Asking clinicians not to use intervention content when treating those in control arm	3
	Providing active intervention within the research project rather than health service	1
	Using a script for contact with control participants during treatment	1
Clinicians not involved in active intervention treating participants in both trial arms	Blinding usual care clinicians	4
	Confining intervention to provision by specialist clinicians	2
Communication between clinicians in different trial arms	Asking clinicians not to share details of the intervention with each other	5
Communication between participants in different trial arms	Holding treatment sessions at different times/in different locations	13
	Staggering the scheduling of data collection appointments / reducing waiting time so that participants do not meet in waiting room	3
	Allocating separate therapists / modes of delivery for individual and group therapies when usual group therapy was shared by participants in both arms	2
	Asking participants not to share contents of intervention with others	2
	Excluding potential participants who know someone else attending screening	2
	Holding separate sessions of existing group treatments for participants in separate trial arms in order to prevent contact	1
	Restricting the release of intervention materials in order to reduce the chance of their being shared with control participants	1
	Recruiting participants in blocks and providing one treatment at a time, with no new participants recruited during the final week of each period in order to maintain separation between trial arms	1
Participants switching clinicians and therefore trial arms	Preventing referrals for add-on care by clinicians who are members of study team	1
	Avoiding transfer of participants between clinicians	1
Participants seeking treatment outside the trial	Informing participants only about the treatment they were allocated to receive (Zelen's design)	8
	Promising the intervention to control participants at the end of follow-up	2
Active treatment is available to some extent within the healthcare system	Making intervention distinct from usual care by adapting one or other	2
	Establishing common treatment for all participants	1
	Excluding institutions that already offer some aspect of the intervention	1

Table 2.5: Summary of trials using both cluster- and participant-level treatment allocation.

Reference	Intervention	Comparator treatment	Population	Sample size
Clarkson et al. (2009)	Dentist consultation targeting patient self-efficacy (using social cognitive theory) and action plans (implementation intention theory)	Routine care	Adults having routine dental check-up	$n = 300$ (individual randomisation) $n = 478$ (cluster randomisation)
Lee and Gay (2011)	Sleep hygiene intervention package for parents (bassinet, white noise machine, night light)	Attention control	Couples and women expecting first child with no history of sleep disorder	$n = 152$ (individual randomisation) $n = 152$ (cluster randomisation)
Marshall et al. (2004)	Systematic assessment of patient's needs and feedback for care coordinator	No feedback to care coordinators	Patients with severe mental disorder being cared for in community	$n = 140$ (individual randomisation) $n = 164$ (cluster randomisation)
Richards et al. (2008)	Collaborative care protocol (coordinated medication support, brief psychological treatment, and enhanced specialist and GP communication)	Usual care management	Adults diagnosed with depression by a GP	$n = 79$ (individual randomisation) $n = 76$ (cluster randomisation)

2.3.7 Trials using both cluster- and participant-level treatment allocation

The results of the review included four trials that used both participant- and cluster-level treatment allocation (Clarkson et al., 2009; Lee and Gay, 2011; Marshall et al., 2004; Richards et al., 2008). The characteristics of these trials are summarised in Table 2.5. The treatments and patient populations are disparate but have in common the use of both cluster and individual randomisation.

Treatment effect estimates and confidence intervals for the four trials are shown in Figure 2.3. The figure shows treatment effects arranged such that greater benefit (or less harm) of treatment is represented by a greater number on the horizontal axis. The figure enables the comparison of the absolute size of treatment effect between participant- and cluster-level allocation to assess the impact of contamination on effect size estimation.

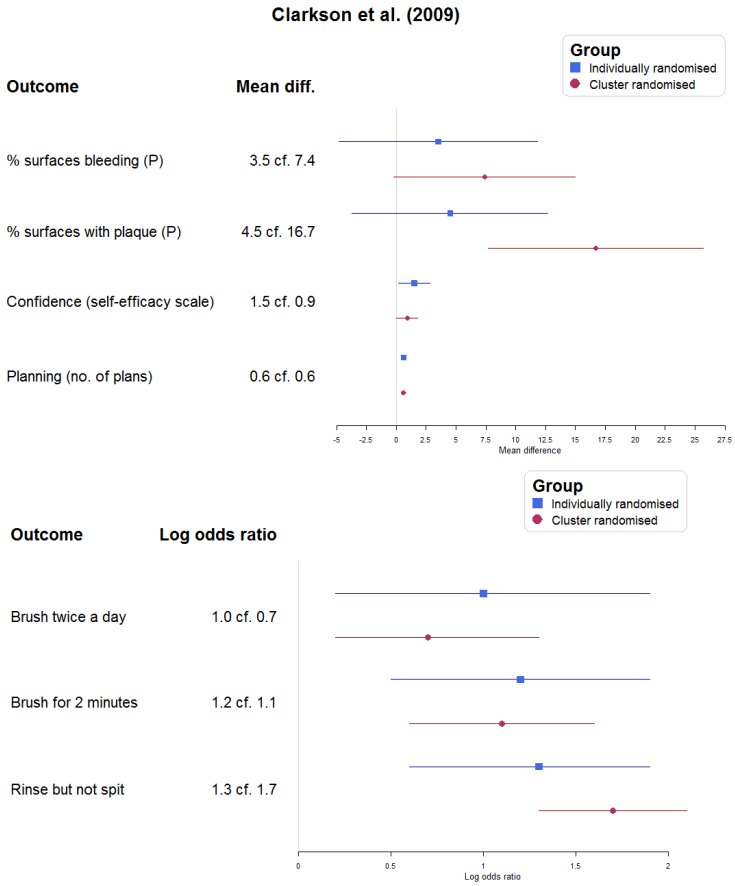
Of the 21 outcomes investigated, just under half of outcomes showed a difference in the anticipated direction, i.e. smaller estimated absolute effect sizes under participant-level random allocation. In particular, an attenuated absolute treatment effect size (lesser distance from the null line in Figure 2.3) was found under participant-level allocation in eight out of 21 outcomes with a tie in one outcome.

2.3.8 Parameters

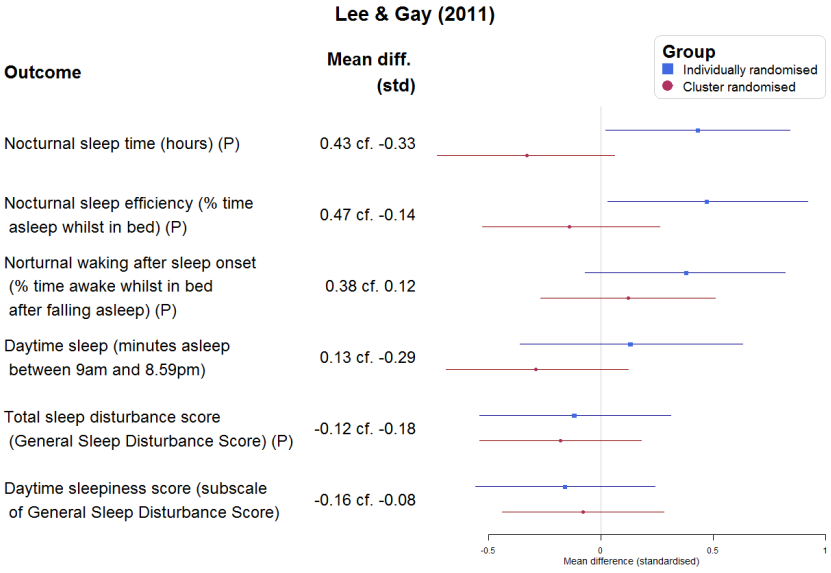
Table 2.6 summarises cluster sizes, intraclass correlation coefficients, and sample sizes for the trials where it was possible to extract this information. The summaries are given for all trials, and for individual randomised and cluster randomised trials separately.

Table 2.6: Summary of clustered data and sample size parameters.

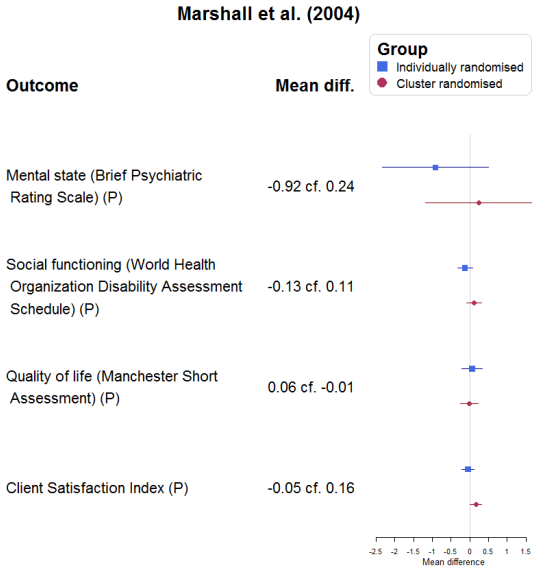
Parameter	**Percentile**	**Individual randomised trials**	**Cluster randomised trials**	**Overall (all trials)**
Cluster sizes	25		6	6
	50		10	10
	75		27	27
ICC	25		0.03	0.03
	50		0.05	0.05
	75		0.09	0.09
Sample size	25	84	153	100
	50	144	266	186
	75	265	556	372



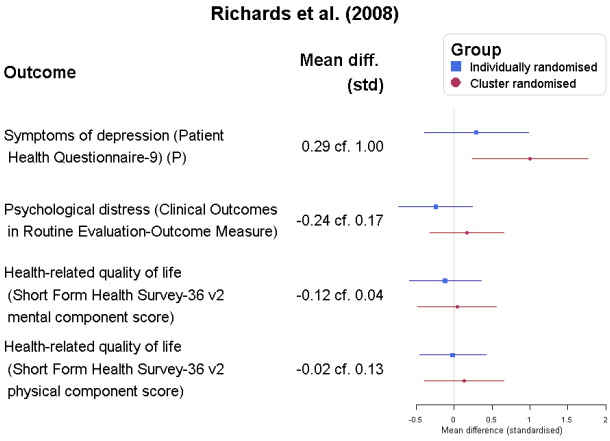
(a) Clarkson et al. (2009). Choice of primary outcomes is based on sample size calculation; estimates are adjusted for baseline measures. Larger (more positive) treatment effects indicate benefit.



(b) Lee and Gay (2011). Estimates were standardised and calculated from summaries of means and SDs (mothers' scores only). Larger (more positive) treatment effects indicate benefit.



(c) Marshall et al. (2004). Estimates used same adjustments as in the trial publication. Larger (more positive) treatment effects indicate benefit.



(d) Richards et al. (2008). Estimates were standardised and calculated from summaries of means and SDs. Larger (more positive) treatment effects indicate benefit.

Figure 2.3: Forest plots for four trials that used both individual- and cluster-level randomisation (P)=primary outcome.

2.4 Discussion and conclusion

2.4.1 Discussion

The review identified 238 trials that described either the processes driving treatment contamination, its quantity, or steps that researchers took to prevent or alleviate the problem in trials of complex interventions in mental health. The principal processes leading to contamination were found to be clinicians being required to treat participants in both treatment and control conditions and communication among clinicians or participants in different trial arms. Typically around one in eight participants in the control arm of a trial were assessed as having received the active intervention. The most common steps that researchers took to prevent or mitigate contamination were the use of cluster randomisation, organising for each clinician to provide only one type of treatment, monitoring treatment receipt, spatially or temporally separating trial arms, and informing participants about only the treatment that they were allocated to receive. There was little evidence of a difference in the magnitude of treatment effects within trials that used both cluster- and participant-level treatment allocation.

The classification of two main processes and three more minor types of contamination was based on the processes that researchers and clinicians described in such trials. The main trial conduct steps that researchers took to minimise contamination were in line with the processes that were found to be driving it. Many researchers attempted to design against contamination by carefully controlling the treatment's delivery. There were no examples of researchers first having evaluated in detail treatment receipt within the control arm.

The small number of trials that measured and reported treatment receipt in the control arm found it to be affecting a minority of the control participants. The distribution of this was similar to the quantity found previously in other areas of medicine such as educational interventions (Keogh-Brown et al., 2007), breast cancer screening (Goel et al., 1998), and cancer trials using Zelen's design (Torgerson, 2001). Thus while there is a lot of concern about contamination, it is not clear that this problem is indeed widespread.

Researchers often used cluster randomisation to prevent treatment contamination, amongst other reasons. While CRCTs can avoid contamination bias they are at risk

of other biases. The set of articles included 93 cluster randomised trials. Assessments of bias suggested that cluster randomised trials were more likely to be affected by bias when considering whether randomisation took place after consent was obtained and after baseline measures were completed, whether outcome assessment was blind, and whether attrition was similar between trial arms. This was consistent with an earlier review of cluster RCTs that were published in three prominent medical journals which found evidence of recruitment and attrition biases (Puffer et al., 2003).

The small number of trials that allocated treatment at both cluster and participant levels did not find any evidence for differences in effect size estimates. The lack of evidence for a link between the level of randomisation and treatment effect size suggested that either the employment of cluster randomisation did not prevent contamination, the anticipated contamination was overstated, or that the use of cluster randomisation led to a similar degree of bias as that caused by contamination in the participant-randomised trials. Overall, the finding was consistent with those of a review of trials of enhanced care in depression (Hahn et al., 2005), and of educational interventions (Keogh-Brown et al., 2007). Similarly to previous reviews, there was considerable heterogeneity between trials identified in this study that used both cluster- and participant-level randomisation. However, the variability here is between trials and not within them because randomisation implies that the subtrials were balanced for every variable except the level at which treatment allocation took place. It is possible that the impacts of contamination and cluster randomisation on bias are dependent on the disease or type of intervention.

There were several limitations to this scoping review. For instance, the processes considered to be driving contamination were often anticipated and then prevented or attenuated by the designers of trials. The drivers of contamination described here therefore partly represent researchers' expectations and not necessarily clinician or participant behaviour. Regarding trial conduct solutions for preventing contamination, the most common method was to recruit groups of clinicians where a given group was responsible for providing a particular treatment. I did not record the methods by which clinicians were allocated to these groups. This is something that would be interesting to consider in future reviews. Finally, it is difficult to draw substantive conclusions regarding the effect of treatment allocation levels on contamination bias from the four trials that used both cluster- and individual-level randomisation. This is due to the small number of such trials that I found and the fact that the review did not search for these trials in particular. There would

be some utility in conducting a systematic review of these trials in future in order to investigate further the effects of allocation level on contamination bias.

2.4.2 Conclusion

This is the most comprehensive review of contamination in mental health trials to date. It is the first to identify the processes leading to contamination and the measures that researchers take in order to minimise the problem. The main limitation is that the trials were heterogeneous in that they represented a large range of illnesses and interventions. With regard to the causes of contamination, it is an assumption that the processes described by authors were the drivers of contamination.

The results of this review suggest that treatment contamination is perceived to be a significant problem in trials of complex interventions in mental health. However, the trials that measured and reported it suggest that the phenomenon is often modest (with a large range). This implies that contamination may not be as large a problem as many researchers and funders perceive. It is therefore not impossible that the use of cluster randomisation in some trials may be leading to efficiency loss for little reason. The seemingly modest extent of contamination points towards the importance of alternative approaches for addressing the problem. The findings of this review have shown that there are many such (conduct) methods that researchers can take to minimise contamination without resorting to cluster randomisation. A caveat to the conclusion about the extent of the problem is that the number of trials in the review that reported contamination was low. The reporting of treatment receipt in the control arm is almost certainly not as commonplace as that of treatment compliance (in the intervention arm). This implies a need for greater measurement and reporting of treatment receipt in the control arm of trials in this field.

I found that modern causal analysis methods, including the techniques developed particularly for addressing contamination (Cuzick et al., 1997; Dunn et al., 2005), are yet to be utilised to deal with contamination bias at the analysis stage. This may be a reflection of the infrequency of measurement of treatment receipt for all participants in the control trial arm. It may also be due to the fact that moderately little research has investigated the use of randomisation-based efficacy estimators in this context, particularly in the field of mental health research. I will return to this topic to summarise and develop such analytical methods in Chapter 4.

Chapter 3

Assessing treatment fidelity and contamination in a cluster randomised controlled trial of motivational interviewing and cognitive behavioural therapy skills in type 2 diabetes.

3.1 Background and aims

The thesis uses the D6 study as a motivating dataset of a trial of a complex intervention in mental health. D6 was a cluster RCT set in primary care evaluating the effectiveness of an intervention combining motivational interviewing (MI) and cognitive behavioural therapy (CBT) skills delivered by practice nurses compared to an attention control which did not include any psychological components. One reason for using cluster randomisation at the level of the primary care nurse was to avoid treatment contamination that was anticipated if a given nurse were asked to provide both control and active treatments. Treatment in the control arm consisted of standard diabetes care, with primary care nurses scheduled to meet participants for the same number of times and same duration as those in the active intervention arm.

There was a strong rationale for conducting a large fidelity assessment. Despite the trial having set out to avoid contamination, there was some anecdotal evidence of delivery of psychological treatment by standard care nurses. This was supported by the treatment fidelity evaluation in the primary assessment of the trial which showed some treatment receipt in the control arm (Ismail et al., 2018). In addition, the lack of evidence of effectiveness warranted a detailed evaluation of treatment fidelity in both arms of the trial. Therefore I set out to construct a measure of treatment receipt for as many participants as possible in the study. The trial provided an opportunity to assess treatment fidelity due to the fact that many treatment sessions had been audio recorded.

The goal was to use individual-level fidelity ratings to assess what level of psychological treatment participants received and then to evaluate whether the treatments delivered to the intervention and control arms represented what was expected given the results of random treatment allocation. From a clinical perspective this treatment fidelity assessment, which was larger in terms of sample size than that reported in the primary assessment of the trial, enabled an examination of whether primary care nurses in the trial could be trained to deliver psychological therapy competently to participants within the active intervention arm. It also allowed an assessment of whether competencies improved over time. This chapter describes the fidelity assessment of the treatments delivered to participants in the two trial arms.

The main aims of this chapter were to measure treatment fidelity in the D6 trial and then to assess treatment contamination and non-compliance. Specifically, the chapter aimed to:

i. Assess adherence in delivering randomised treatment within each trial arm,

ii. Determine to what extent the intervention and control treatments represented high fidelity psychological treatment and standard diabetes care respectively.

Treatment fidelity was assessed for a large subsample of participants and therefore allowed an assessment of treatment receipt that was more detailed than that of the primary assessment of the trial (Ismail et al., 2018). Therefore, the clinical aims of this chapter were to:

i. Assess whether D6 nurses achieved competencies in psychological therapy delivery at the end of the training period,

ii. Describe differences in competency between end of training and during the delivery of intervention,

iii. Compare the levels of receipt of psychological treatment (MI and CBT skills) between the active intervention and control arms.

This chapter will describe the sampling procedure and the implementation of the fidelity ratings. The sampling was at the level of the participant as opposed to sampling a particular number of sessions for each nurse, as done in the primary fidelity assessment (Ismail et al., 2018). I performed the sampling process and then oversaw the preparation of the recordings (finding and sorting the tapes) that was done by an administrator from the trial. The ratings were then performed by two clinical psychologists who had experience in the scales that were used to rate treatment fidelity. Data that were collected soon after the end of training (intervention arm) will be used to assess nurse competency. Data collected as part of the doctorate (on treatment fidelity during delivery of treatment) will be used to assess contamination and non-compliance. These data will also be used to evaluate change in nurse competency (intervention arm) and to assess differences between the trial arms. The data will later be used to construct measures of treatment receipt for the efficacy analysis of the trial (Chapter 7).

The work presented here has been submitted for publication as an article to BMC Family Practice.

3.2 Methods

3.2.1 The training programme

The training programme for nurses in the D6 intervention arm was developed and delivered by an experienced clinical psychologist using both didactic and practicum strategies. Nurses were trained in six MI/CBT skills: active listening, managing resistance, directing change, supporting self-efficacy, addressing health beliefs, and shaping behaviours. The initial interactive training workshops were conducted over twelve 3-hourly sessions and the nurses were given a manual handbook for ongoing reference and for future replication. The focus was on increasing patients' motivation to improve their diabetes control and then collaboratively addressing key self-care behaviours such as medication adherence, blood sugar testing, physical activity, and dietary changes.

3.2.2 Techniques taught in MI and CBT

MI is a collaborative, person-centred approach to working with people in order to elicit and strengthen their motivation and commitment to change (Miller and Rollnick, 2002). It has been found to be more effective than traditional advice-giving in the treatment of a range of behavioural problems and diseases, including diabetes (Whittemore et al., 2003). CBT has been found to be effective at improving adjustment to diagnosis and self-management of diabetes (Ismail et al., 2004). It aims to achieve this by helping people to identify and restructure unhelpful cognitions, teaching behavioural strategies, and supporting people to develop helpful coping strategies.

3.2.3 Clinical supervision

Nurses in the intervention group attended monthly supervision with the trial psychologist either in person at monthly group sessions or over the telephone if they were not able to attend throughout the delivery of the intervention. E-mail support was also offered for individual cases.

3.2.4 Assessment of treatment fidelity and competency

All nurses who participated in the D6 study were required by protocol to record their treatment consultations with participants digitally. A sample of recordings from nurses in the intervention arm from shortly after the end of training was used to assess competency. Another sample of recordings from both trial arms that was representative of participants' treatment receipt was selected in order to assess fidelity.

The definition, assessment, and difficulties of addressing treatment fidelity in research studies have been extensively discussed elsewhere in the literature (Madson and Campbell, 2006; Rakovshik and McManus, 2010; Fairburn and Cooper, 2011). A definition that is consistently used, and will be used for the purpose of this chapter, is that fidelity comprises both adherence and competence (Fairburn and Cooper, 2011). Adherence refers to whether the appropriate procedures were followed for that clinical intervention whereas competence refers to whether these procedures were implemented to an adequate level.

The Motivational Interviewing Treatment Integrity (MITI) Scale, version 3.1.1 (Moyers et al., 2005, 2010), was used to measure competence and skills used in both groups of nurses. A Global Spirit score is intended to capture the overall demonstration of

MI principles, and a Global Empathy score is intended to capture the extent to which the clinician understands, or attempts to understand the patient's perspective. Further measures of clinician behaviours include the use of simple reflections, complex reflections, open questions, and closed-ended questions. Scores are also calculated for MI adherent and non-adherent counselling behaviours. The possible ranges and threshold levels for subscales (as specified by the scale's authors) are given in Table 3.1.

The Behaviour Change Counselling Index (BECCI) (Lane et al., 2005) was designed to assist trainers in assessing a clinician's competence in using behaviour change counselling in consultations. It was included here in order to assess nurses' competence in eliciting patients' thoughts and cognitions, therefore addressing the CBT element of the intervention. BECCI comprises 11 items which are scored from zero to four (0="action carried out not at all"; 1="minimally"; 2="to some extent"; 3="a good deal"; 4="a great deal"). The mean of these is used as the overall Practitioner Score.

Table 3.1: Minima, maxima, and thresholds for MITI and BECCI scales (Moyers et al., 2005, 2010; Lane et al., 2005).

Scale	**Minimum (lowest score)**	**Maximum (highest score)**	**"Beginning proficiency"**	**"Competency"**
MITI summary scores				
Global Clinician Ratings (spirit and empathy)	1	5	Average of 3.5	Average of 4
Reflection to Question Ratio	0	–	1	2
Percent Open Questions	0	100	50%	70%
Percent Complex Reflections	0	100	40%	50%
Percent MI-Adherent	0	100	90%	100%
BECCI summary score				
Practitioner Score	0	4	–	–

This chapter utilises three datasets: a "nurse competency" sample and two "fidelity assessment" samples. The nurse competency sample, which was collected previously as part of the D6 trial, included one tape recording for each intervention nurse (11 nurses). The first fidelity assessment sample (69 recordings from 21 nurses; sampling at level of nurse) was used for quantifying the reliability of the ratings made by the clinical psychologists working on this research. These data had already been collected as part of

the primary assessment of D6. The second fidelity assessment sample was larger (266 recordings from 151 patients and 17 nurses; sampling at level of participant) and was used for the fidelity assessment, which is the main focus of this chapter. These fidelity ratings were collected as part of this doctoral research and the goal was to use these data to construct a measure of treatment receipt for the efficacy analysis of the trial (Chapter 7).

3.2.5 Nurse competency assessment

The nurse trainer, who was MITI trained, assessed post-training adherence and competency of all nurses in the intervention group using the MITI and BECCI rating scales. One tape recording of a treatment consultation was submitted by each nurse soon after the end of training and then rated on each of the two scales. Nurses were rated as not MI adherent if MITI MI-Adherence was lower than 90% (the "Beginning proficiency" threshold, see Table 3.1) and MITI Empathy was lower than 3 (which is defined as representing modest success of clinician trying to understand the patient's perspective (Moyers et al., 2010)). These subscales were chosen because MI-Adherence and Empathy have been shown to be predictive of treatment success (Apodaca and Longabaugh, 2009; Moyers and Miller, 2013). The "Beginning proficiency" and "Competency" thresholds in the MITI manual (Table 3.1) were considered *post hoc* too high in the context of D6, where consultations included clinical communications that would not be part of a standard MI consultation (for example a physical examination, prescribing, and checking adherence). Any nurses rated as not MI adherent were given extra training and then reassessed. Nurses who were judged to be adherent but who did not meet the higher MITI threshold levels were expected to continue to improve with extra supervision.

3.2.6 Sampling for inter-rater reliability assessment

For the first fidelity assessment (collected previously as part of the primary assessment of D6) a researcher assessed every tape recording and removed duplicates and recordings where session number could not be identified. Of the tape recordings that were from treatment sessions two to four, and where there was a recording of a treatment session that lasted 20 minutes or more, stratified probability sampling was used to select three recordings from each nurse. Within each nurse stratum, the first tape recording was chosen at random and the second recording was then chosen at random after removing recordings from the previously-selected individual and session from the sample set. The

same technique was used to sample the third recording.

The sample comprised 69 tape recordings (representing 3.4% of the total number of all treatment sessions, and 4.0% of sessions where a recording had been made). A 20-minute window in the middle of the recording was rated using the MITI (by raters A and B). Of this sample, 32 recordings were rated using the BECCI by raters B and C. Recordings in this subsample featured in both the reliability assessment and fidelity assessment (described in next section). Rater C listened to and coded a 20-minute window in the middle of the recording whilst rater B assessed the entire recording (raters B and C's assessments were originally intended for different purposes). Raters received suitable training for whichever scale they used and were blind to treatment allocation. This sample was used in order to check the inter-rater reliability of raters who assessed recordings in the fidelity study.

3.2.7 Sampling for fidelity assessment

The sampling procedure for the second fidelity assessment sample (data collected as part of this doctoral research) selected tape recordings from participants who had at least one recording from sessions two, three, and four, and where treatment centre was identifiable (there was no minimum duration of session length). This set included 353 recordings from 154 participants (31 participants with one recording; 47 with two; and 76 with three). Random sampling stratified by participant was used to select two recordings from each of the participants with all three recordings. If only one or two recordings were available for a given participant then these were chosen for subsequent fidelity assessment.

The sample included 266 usable tape recordings (127 recordings in intervention arm) from 17 nurses' consultations with 151 participants and 11 recordings where the conversation could not be heard. The usable recordings represented 13.1% of all treatment sessions and 15.4% of sessions where a recording was made. The whole duration of each recording was rated using the MITI (rater A) and BECCI (rater B). Raters were blind to treatment allocation.

3.2.8 Statistical analysis

Statistical analyses were conducted using Stata version 14. In order to assess inter-rater reliability for the MITI global scores and BECCI Practitioner Scores, intra-class correlation

coefficients (ICCs) were estimated using a mixed model. The model included a fixed effect for rater, a random effect for tape recording, and a random effect for primary care nurse in order to account for clustering. It assessed consistency between individual ratings by estimating ICCs at the participant-within-nurse level. The MITI global scores and BECCI Practitioner Score were summarized within the intervention arm shortly after the end of training and during delivery of intervention. Mixed effects regression models with random effects for primary care nurse and participant or Somers' D tests with sampling from the highest level of the cluster structure (i.e. primary care nurse) were used to compare the fidelity of the psychological therapy delivery between participants in the two trial arms.

3.3 Results

3.3.1 Nurse and patient sample characteristics

Twenty-three primary care nurses participated in the trial, with 11 randomised to the intervention arm, and 12 to control. They were all female, with a mean age of 48 (SD 8.5) years. Fourteen (61%) of the nurses were white, six (26%) black, and 3 (13%) Asian or other ethnicity.

In terms of previous training in psychological therapies, nine had no previous experience (4 intervention, 5 control), two had completed a module as part of a degree course (1 intervention, 1 control), two had completed some training in MI as part of a smoking cessation course (1 intervention, 1 control), two had undertaken one day or less of MI training (1 intervention, 1 control), one had completed some MI training as part of the Co-Creating Health Programme (intervention), and one had some experience as part of a nursing qualification (intervention). Data on previous training were not available for six nurses.

The participant sample from which the tape recordings were drawn (treatment fidelity assessment sample) included 151 adults with T2D (45% of the total number of participants who entered into the trial), of whom 74 (49%) were in the psychological treatment trial arm. Mean age was 59.4 (SD 11.1) years and 77 (51%) were female. Sixty-eight (45%) were white, 60 (40%) black, 13 (9%) Asian, and 10 (7%) of another ethnicity. Median duration of diabetes was 9 (IQR 6-13) years and mean pre-intervention glycated haemoglobin was 80.1 mmol/mol (SD 18.9).

3.3.2 Nurse competency

The trial manager assessed post-training treatment adherence and competency using the MITI and BECCI rating scales. Mean MITI and BECCI competency scores post-training are presented in Table 3.2. Each nurse's score was compared against MI-Adherence (threshold of 90%) and Empathy (threshold of 3). One nurse was not considered MI adherent post training (using MITI MI-Adherence and Empathy subscales) and therefore was given extra training by the clinical psychologist. Upon reassessment she was deemed MI adherent in the therapy.

Table 3.2: Summary of competency scores assessed after training.

Domain	Post-training score
Global Spirit (mean; SD)	3.42 (0.67)
Global Empathy (mean; SD)	4.09 (1.04)
Reflection-to-Question Ratio (median; IQR)	0.67 (0.45-0.82)
Percent Open Questions (median; IQR)	45.5 (25.0-72.2)
Percent Complex Reflections (median; IQR)	9.1 (0-28.6)
Percent MI-Adherent (median; IQR)	86.2 (76.9-100)
BECCI (mean; SD)	2.78 (0.50)

3.3.3 Inter-rater reliability

Estimates of intraclass correlation coefficients for the global MITI scores and BECCI Practitioner Score are reported in Table 3.3. These estimates suggested that inter-rater reliability was good (between 0.60 and 0.74) or excellent (>0.75) for both scales, according to previously defined thresholds (Cicchetti, 1994). Reliability was greater for MITI, where all ratings were for the 20-minute section in the middle of each recording, compared to BECCI, where one coder rated 20-minute windows and another rated the full duration of recordings.

Table 3.3: Inter-rater reliability for MITI global scores and BECCI Practitioner Score.

Domain	ICC	95% confidence interval
MITI Global Spirit	0.89	0.83–0.93
MITI Global Empathy	0.91	0.86–0.94
BECCI Practitioner Score	0.71	0.52–0.85

3.3.4 Fidelity analysis

MITI domain scores summarised by trial arm along with the results of the mixed model or Somers' D tests comparing trial arms are given in Table 3.4. Estimated standardised mean differences for the MITI global scores were 1.11 (Spirit) and 0.83 (Empathy). There was strong evidence of group differences in favour of the intervention for the global scores of Spirit and Empathy, the percentage of questions that were open, and of percentage of sessions that were MI adherent. There was no evidence of a group difference in percentage of reflections that were complex or the reflection-to-question ratio.

Table 3.4: MITI summary scores during treatment delivery by treatment allocation group.

MITI Domain	**Usual care group (mean; SD)**	**Intervention group (mean; SD)**	**z-test (from mixed model)**	**95% confidence interval for mean difference**
Global Spirit	2.63 (1.12)	4.03 (1.05)	$z = 4.50$; $p < .001$	0.81–2.06
Global Empathy	3.40 (0.98)	4.23 (0.89)	$z = 4.55$; $p < .001$	0.49–1.23
	Usual care group (median; IQR)	**Intervention group (median; IQR)**	**z-test (from Somers' D)**	
Reflection-to-Question Ratio	0.50 (0.33–0.71)	0.44 (0.32–0.61)	$z =$ -0.55; $p = .58$	
Percent Open Questions	23.1 (13.3–37.5)	46.5 (33.3–57.1)	$z = 4.17$, $p < .001$	
Percent Complex Reflections	55.6 (41.9–71.4)	53.8 (40.0–71.4)	$z = 0.12$, $p = .90$	
Percent MI-Adherent	21.4 (10.0–35.0)	63.4 (33.3–83.3)	$z = 3.68$; $p < .001$	

Numbers and proportions of sessions in the intervention arm that were rated as above MITI's "Beginning proficiency" and "Competency" thresholds for each domain are summarised in Table 3.5 (Moyers et al., 2005). This table summarises how many treatment sessions were assessed as meeting these thresholds within each of the trial arms.

Mean BECCI Practitioner Score in the control arm was 1.07 (SD 0.48) and in the intervention arm was 1.42 (SD 0.51). A z-test from a mixed effects model showed a significant

Table 3.5: Numbers and proportions of sessions (percentages are within trial arms) rated as above MITI's "Beginning proficiency" and "Competency" thresholds for domains by treatment allocation group.

MITI Domain	Beginning proficiency		Competency	
	Number above threshold in standard care group (%)	Number above threshold in D6 intervention group (%)	Number above threshold in standard care group (%)	Number above threshold in D6 intervention group (%)
Global Spirit	34 (24.5)	92 (72.4)	30 (21.6)	88 (69.3)
Global Empathy	71 (51.1)	103 (81.1)	71 (51.1)	103 (81.1)
Reflection-to-Question Ratio	17 (12.2)	9 (7.1)	4 (2.9)	0 (0)
Percent Open Questions	13 (9.4)	54 (42.5)	5 (3.6)	9 (7.1)
Percent Complex Reflections	106 (76.3)	98 (77.2)	87 (62.6)	78 (61.4)
Percent MI-Adherent	1 (0.7)	26 (20.5)	1 (0.7)	25 (19.7)

difference in the BECCI Practitioner Scores between the treatment arms ($z = 3.22$, $p < .01$, 95% CI 0.15-0.62). The estimated standardised mean difference was 0.75.

3.4 Discussion and conclusion

3.4.1 Discussion

This chapter describes the assessment of the delivery of a nurse-led psychological therapy in the context of a cluster RCT aimed at improving persistent suboptimal glycaemic control in people with T2D. Treatment fidelity and contamination were evaluated by measuring and comparing levels of MI and CBT skills in the two trial arms. At the end of training, nurses in the intervention group were considered competent in D6 skills at a basic level (according to "Beginning proficiency" thresholds) and it appears that there was improvement in some MI skills during delivery of the intervention. For example, MITI Global Spirit and the proportion of reflections that were complex improved. The active intervention delivered to trial participants was statistically superior in Spirit and Empathy, open questions, MI-Adherence, and behaviour change scores compared to usual care. There were no group differences in the proportion of complex reflections or the

reflection-to-question ratio. In clinical terms, the differences between the trial arms were smaller than expected. The levels of treatment fidelity suggested that some participants in the psychotherapy arm did not receive high fidelity treatment, whilst some in the attention control arm received aspects of the psychological intervention. These findings highlight the difference between statistical and clinical significance. Statistically significant differences primarily reflected the fact that this was an unusually large assessment of treatment fidelity in an RCT. However, many of these significant differences in MITI and BECCI scores between trial arms were clinically small. For example, the difference in the BECCI Practitioner Score was less than half a point (on a scale of zero to four) and was significant at the 1% level. The lack of clinical significance suggestested that the estimate of effectiveness of the D6 intervention (ITT analysis) might have been affected by both non-compliance in the active arm and contamination in the control arm.

In the active intervention arm, findings were partly consistent with the practice of MI, where the clinician collaborates with, supports, and allows the patient to take control of the need for change by listening empathically and using open-ended questions. This was demonstrated by high levels of Spirit and Empathy and a clear majority of treatment sessions being MI-Adherent. The superiority of MI-Adherence and Empathy when comparing the trial arms was particularly important as these have been shown to be predictive of treatment success (Apodaca and Longabaugh, 2009; Moyers and Miller, 2013). However, there were some challenges in providing high fidelity psychotherapy. Specifically, approximately only half of reflections were complex, a similar proportion of questions were open, the ratio of reflections to questions was slightly lower in the intervention group compared to control, and the level of achieved behaviour change fidelity (from the BECCI) was rated between “minimal” and “to some extent”.

There were a number of possible reasons why nurses may not have exceeded MITI’s “Beginning proficiency” levels. The most apparent of these is that the nurses did not self-select to take part in D6. All primary care surgeries meeting the eligibility criteria in the five boroughs were invited to participate. Of those that agreed, the surgery allocated a nurse to take part in the study. Some nurses were more enthusiastic about participation than others. It is also possible that the skills that showed the lower fidelity levels reflected particular aspects of MI or CBT that are difficult to teach to clinicians who are not specialists in psychological treatment. An interview study with the nurses suggested that not all may be suited to the acquisition of psychological skills (Graves et al., 2016).

For example, nurses expressed concern about over-stepping their professional roles, feeling that it was inappropriate for them to deliver specialist psychological intervention and described feeling under pressure to participate in the research. Some felt under-supported by their primary care surgery and others resented the extra workload as a result of participating in the trial. Although the surgeries were remunerated for participation, the trial did not provide direct individual financial compensation. One solution to this problem may be to assess inherent competencies prior to training, enabling a process of selection whereby the most suitable nurses are recruited. This is a similar idea to that put forward in an assessment of treatment fidelity of nurse-led MI in pain rehabilitation, where the authors suggested that more rigour was necessary in the selection of MI counsellors (Mertens et al., 2016). It is not currently possible to distinguish whether D6 nurses possessed existing psychological skills, which were not especially built upon, or whether they learned skills to a basic level but then failed to improve materially upon them.

In the attention control arm, the moderate levels of Spirit and Empathy of MI, the ratio of reflections to questions, which was slightly higher than in the psychological treatment arm, and the fact that just over half of reflections were complex showed clear evidence that there was delivery of MI. On the other hand, the behavioural change index summary score was low in this trial arm. The evidence of delivery of active intervention in the control arm was surprising given that the trial was designed to avoid this. It was hypothesised before starting the trial that contamination might have been expected if a given clinician were to be trained in the delivery of the psychological treatment and then treat participants in both trial arms. The expectation was that if this were to happen, clinicians would have introduced elements of the active intervention to participants in the attention control arm. Cluster randomisation was used in part to avoid contamination occurring in this way. The contamination that took place despite this design must have been due to other reasons. For instance, some primary care nurses already possessed skills that were consistent with psychological treatment. Two control nurses are known to have had experience of MI before the trial: one had received brief training in it and one had applied it to smoking cessation. Other reasons include the impact of giving extra time to deliver standard care as part of the attention control design; finally, participation in the trial may itself have induced nurses to provide a slightly different type of standard care.

The primary analysis of D6 included a fidelity assessment of a small sample of therapy session recordings (n=69) in both treatment groups, using both the MITI and the BECCI (Ismail et al., 2018). The researchers sampled three tape recordings from each nurse and rated only a 20-minute window in the middle of each recording. The findings showed similar trends to those reported here, but the trial arm differences were estimated to be smaller and had larger standard errors. This demonstrates a benefit of rating treatment fidelity for participants (ideally a large sample or all of them) rather than clinicians. For a trial investigating efficacy, a full assessment of treatment fidelity with a representative sample is needed in order to use a measure of treatment receipt to estimate this target effect. The limitations of a full assessment are its labour-intensive nature and the increased costs of employing trained raters. Costs may come down with developments in machine learning and automated fidelity evaluation.

Future research should ensure that therapists in the experimental arm of a complex psychological intervention are at higher level of competency in order to make more valid comparisons with the control.

3.4.2 Conclusion

In summary, the results indicate that the intervention did not represent the highest level of psychotherapy fidelity. In addition there seemed to be some contamination of the control arm as those allocated to receive usual care appeared to receive some components of the intervention. On the basis of these findings it appears that the trial would benefit from an efficacy assessment in addition to the effectiveness evaluation (ITT analysis) that has already been carried out. This will be the subject of Chapter 7. To enable this, I first need to summarise and develop statistical methods for estimating efficacy. This will be the subject of the next chapter.

Chapter 4

Efficacy estimators allowing for non-adherence in trials of complex interventions

4.1 Background and aims

This chapter addresses the first part of the secondary research objective, which was to summarise and develop estimators of efficacy in a randomised trial with non-adherence. This work focuses on trials that contrast an active intervention with a control condition, where departures from randomised treatment lead to receipt of some or all of the comparator condition. The work relates to trials with continuous outcomes.

The perfect randomised trial comparing a new treatment (intervention arm) with a control treatment (control arm) should use random treatment allocation, be double blind, feature full adherence with randomly allocated treatment, and collect outcome data from every trial participant. These features are important because they guard against bias and therefore enable a study to provide a valid answer to the research question. Judged by such a standard, trials in mental health are often far from ideal. It may be impossible to blind clinicians or participants to treatment (especially for psychotherapies), patients may not attend all or any treatment sessions, and participants may be lost to follow-up before the end of data collection. This has led to recent development in methods that account for the challenges of non-adherence with allocated treatments and missing data in particular. Non-adherence is used here as an umbrella term covering contamination (receipt of intervention) in the control arm and non-compliance (non-

receipt of intervention) in the active intervention arm of a trial. When compliance is described as partial, this implies that some active intervention participants did not receive treatment.

The results of the scoping review of methods for addressing contamination (Chapter 2) suggested a scarcity of applications of analytical approaches in this context. It appears that methods from the causal inference literature are not being utilised by researchers to address the problem of contamination in trials of complex interventions in mental health. Thus, this chapter will explore whether such methods can be employed when adherence with randomly allocated treatment has been measured for each trial participant. The adherence measure could be either a binary treatment receipt measure or a continuous treatment dose measure.

This chapter aims to:

1. Review existing approaches for addressing non-compliance (in the active arm),
2. Extend these approaches to provide novel efficacy estimators that allow for contamination in the control arm,
3. Tackle some of the associated complications experienced in trials of complex intervention in mental health, namely clustering of outcome data and missing values in the outcome variable.

4.2 Estimands

Clinical trials aim to evaluate causal treatment effects. My focus in this thesis is on estimating the efficacy of treatments. Before proceeding to review ways of estimating efficacy, I first wish to define the target population characteristic that I want to estimate. In other words, I wish to define causal estimands. The importance of defining the quantity targeted by an estimator has recently been emphasized – see Addendum E9 trial regulations (European Medicines Agency, 2018). In order to define these estimands, I will begin by introducing the concept of potential outcomes.

4.2.1 Rubin causal model

Potential outcomes were introduced by Neyman (1923), originally in the context of randomised experiments. The use of these provides the main approach to causal estimation

in the *Rubin causal model* (coined by Holland, 1986), which gives the framework for investigation of cause and effect (Rubin, 1974; Rubin and Little, 2002).

4.2.2 Potential outcomes

In a clinical trial, certain variables can be directly observed: which treatment a participant is allocated to, baseline covariates, which treatment they receive, and outcome on a particular scale. The observed variables, their notation, and their levels are described in Table 4.1.

Table 4.1: Notation for observed variables.

Notation	Description	Levels
C	Manifest compliance	0 for observed non-complier 1 for observed complier Missing for those in control arm
K	Contamination	0 for observed non-contaminator 1 for observed contaminator Missing for those in active arm
N	Response (non-missingness)	0 for missing 1 for non-missing
R	Binary indicator for random treatment allocation	0 for assignment to control 1 for assignment to active intervention
T	Treatment receipt	0 for receipt of control treatment 1 for receipt of active intervention
D	Dose of treatment	
X	Baseline covariate	
Y	Outcome	
Z	Baseline variable serving as an instrument (to be defined later)	

The relationship between two variables is said to be causal when outcome differs under the presence of exposure compared to its absence (Hernán and Robins, 2018). For example, if a participant's outcome would be improved under treatment compared to what it would be under no treatment then it can be said that this treatment has a causal effect for that participant. This concept leads to a need for mathematical notation by which outcome can be expressed under different levels of exposure. A participant's possible outcome under a particular level of treatment receipt can be expressed as:

$$Y_i(T = t) = Y_i(t)$$

This is known as a *potential outcome*, or *counterfactual*. $Y(t)$ represents potential outcome under treatment receipt ($t \in 0,1$), and $Y(r)$ is potential outcome under offer of treatment ($r \in 0,1$). Using potential outcome notation, the *causal effect of treatment receipt for individual i* is:

$$\Delta_{T,i} := Y_i(T=1) - Y_i(T=0)$$

The aggregate of these individual effects of treatment on potential outcome among a population is referred to as the ATE:

$$\text{ATE} := \text{E}[\Delta_{T,i}] = \text{E}[Y_i(T=1) - Y_i(T=0)]$$

This is a measure of efficacy. Another efficacy estimand is the average treatment effect on the treated (ATT):

$$\text{ATT} := \text{E}[\Delta_{T,i}|T=1] = \text{E}[(Y_i(T=1) - Y_i(T=0))|T=1]$$

The *causal effect of treatment offer for individual i*, or individual randomisation effect (IRE), is:

$$\Delta_{R,i} := Y_i(R=1) - Y_i(R=0)$$

The aggregate of these causal effects among a population of individuals is known as the average causal effect (ACE):

$$\text{ACE} := \text{E}[\Delta_{R,i}] = \text{E}[Y_i(R=1) - Y_i(R=0)]$$

This is a measure of effectiveness. If adherence were full then T and R would be the same and therefore ACE would be equivalent to ATE. Because the average difference is equal to the difference in averages, the ACE can be rearranged as:

$$\text{ACE} = \text{E}[Y_i(R=1)] - \text{E}[Y_i(R=0)]$$

A potential outcome can also refer to treatment receipt under levels of treatment offer: $T(r)$ is potential treatment receipt under offer of treatment and $D(r)$ is potential dose of treatment under treatment offer ($r \in 0,1$). The pair $T(0)$ and $T(1)$ is used to divide a population into subpopulations, or *principal strata*, that are defined by their potential

treatment receipt under treatment and control. It is not possible to observe both $T(0)$ and $T(1)$ for a given participant and therefore class membership is a latent variable. See Table 4.2 for a list of these potential compliance types, with conventional nomenclature (Angrist et al., 1996). The pair $D(1)$ and $D(0)$ describes subsections of the population defined by doses of active treatment (e.g. number of sessions) they would take were they offered the control or the active intervention.

Table 4.2: Principal strata.

$T_i(R{=}0)$	$T_i(R{=}1)$	Latent compliance class
0	1	Complier
0	0	Never taker
1	1	Always taker
1	0	Defier

The causal effect of treatment defined so far (ATE) refers to the whole target population. The latent compliance classes, whose construction was described earlier and which were listed in Table 4.2, can be used to define a LATE within a subpopulation of would-be compliers. The estimand is known as CACE and is defined as follows:

$$\text{CACE} := \text{E}[\Delta_{R,i} | T_i(1) - T_i(0) = 1] \tag{4.1}$$

If treatment receipt is continuous instead of binary, then the average causal effect must be defined slightly differently. The ATE is now dependent on the values d_0 and d_1 that treatment receipt takes (e.g. number of sessions attended) under the two treatment offers:

$$\text{ACE}_{\text{d}_1,\text{d}_0} := \text{E}[\Delta_{R,i} | D_i(0) = d_0 \geq 0, D_i(1) = d_1 > D_i(0) = d_0]$$

This is the causal effect of treatment amongst a subpopulation who would receive some dose of treatment under its offer and some dose under offer of control, where dose under offer of treatment is greater than dose under offer of control. I refer to this subpopulation as *dose compliers*. This generalises the latent compliers in the binary treatment receipt case to the continuous treatment receipt case.

The next section will address how these estimands can be estimated in a sample.

4.2.3 Identification assumptions of Rubin causal model

It is apparent from the definition of the causal effect for an individual that this cannot be observed because a participant can only receive one treatment (active or control) at a given moment. This is known as the *fundamental problem of causal inference* and has led to the consideration of causal estimation as a missing data problem (Rubin, 2005). The solution to this problem is to take aggregates of outcome under the levels of treatment offer and then find the difference between the averages of these.

There are three main assumptions necessary for the identification of causal effects within the potential outcome framework. First, it is assumed that there is *no interference*, which is to assume that one person's potential outcome is unaffected by another's treatment receipt ($Y_i(t) \perp T_j$, $i \neq j$; Cox, 1958). Second, the assumption of *consistency* assumes that the potential outcome is precisely the same as the observed outcome when exposure takes a particular level; in effect this assumption is that the treatment is well defined. It can be expressed mathematically as:

$$Y_i(t) = Y_i, \text{if } T = t$$

The assumptions of no interference and consistency are together sometimes known as the *stable unit treatment value assumption (SUTVA)* (Rubin, 1980). Third, it is assumed that the distributions of potential outcomes are *exchangeable* between groups defined by level of exposure. There are three levels of exchangeability that are relevant to causal estimation in randomised trials. From strongest to weakest and using similar terminology and notation to Hernán and Robins (2018), these are:

1. *Full exchangeability*: for $t = 0, 1$, $\{Y_i(T = 0), Y_i(T = 1)\} \perp T_i$
2. *Standard exchangeability*: for $t = 0, 1$, $Y_i(T = t) \perp T_i$
3. *Mean exchangeability*: for $t = 0, 1$, $\mathrm{E}[Y_i(T = t)|T_i = 0] = \mathrm{E}[Y_i(T = t)|T_i = 1]$

Full exchangeability implies that the joint distribution of potential outcomes $Y(1)$, $Y(0)$ is independent of T. Standard exchangeability, which here is founded on treatment receipt being dichotomous, assumes that the marginal distributions of the potential outcomes are independent of T. This implies mean exchangeability, which assumes that the expectations of potential outcomes are independent of T. Mean exchangeability does not imply standard exchangeability because it makes no assumption about other

parameters of the distributions of the potential outcomes. Under ideal circumstances (full adherence with protocol and no missing data), randomisation ensures that all three of these assumptions are true because predictors of outcome are equally distributed between treatment groups.

Another way of expressing this is to say that potential outcomes are independent of the mechanism of treatment assignment. Simply put, it assumes no unmeasured confounding between T and $Y(t)$. This untestable assumption is easily defensible in RCTs without non-compliance and contamination, where assignment is random and therefore cannot be associated with any measured or unmeasured pre-randomisation variable, but is more problematic in observational research. It is important to note that the assumption $Y(t) \perp T$ is very different to $Y \perp T$. The first assumes that groups defined by level of treatment receipt would have the same outcome had they received the same treatment, whilst the second assumes that treatment receipt has no causal effect on outcome (Hernán and Robins, 2018).

4.3 Traditional estimation approaches

I first review commonly used estimators in trials, clarifying which estimands they are targeting and discussing possible biases.

4.3.1 Estimating effectiveness: intention-to-treat estimator

The *ITT estimator* estimates the difference in outcome between those allocated to active intervention and those allotted to control ($\mathrm{E}[(Y_i|R_i = 1) - (Y_i|R_i = 0)]$). For random trial samples and under random treatment allocation, an unbiased estimator of the ITT estimator is:

$$\widehat{\text{ITT effect}} := \bar{Y}_1^{(R)} - \bar{Y}_0^{(R)} = \left(\frac{\widehat{\mathrm{Cov}(R_i, Y_i)}}{\mathrm{Var}(R_i)} \right)$$

where $\bar{Y}_1^{(R)} = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_i$; $\bar{Y}_0^{(R)} = \frac{1}{n_0} \sum_{i=1}^{n_0} Y_i$, and n_1 and n_0 are the numbers of active intervention and control participants respectively. This estimator, which will be referred to as **E-ITT**, targets ACE and estimates the effectiveness of treatment. It is unbiased under mean exchangeability which is ensured by randomisation in a trial. The International Conference on Harmonisation (1998) introduced the *ITT principle* which states that this estimator is the best assessment of a treatment, with analysis based on planned treatment

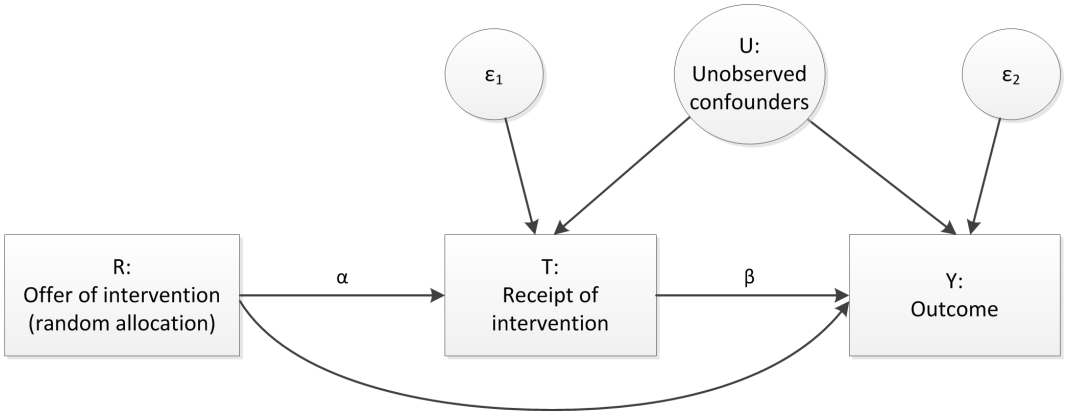


Figure 4.1: Structural equation model illustrating the causal effect of intervention allocation on outcome. Square boxes indicate manifest variables; circles are latent variables.

regimen (the result of random assignment) rather than treatment receipt. In practical terms this means that "subjects allocated to a treatment group should be followed up, assessed and analysed as members of that group irrespective of their compliance to the planned course of treatment" (p.33). The primary analysis of a superiority trial is normally always based on this approach because it is a statistically valid approach for effectiveness (Pocock, 2013) and conservative for efficacy (i.e. estimate of efficacy is biased towards the null; Hernán and Hernández-Díaz, 2012).

4.3.2 Estimating efficacy: as treated and per protocol

Two simple approaches for estimating treatment efficacy are the use of as-treated and per protocol (also known as on treatment) estimators. The *as-treated estimator* estimates efficacy according to the treatment participants received, irrespective of randomisation status, in an attempt to estimate ATE. This is parameter β in Figure 4.1. The as-treated estimator for a continuous outcome is:

$$\widehat{\text{As treated effect}} := \bar{Y}_1^{(T)} - \bar{Y}_0^{(T)} = \left(\frac{\widehat{\text{Cov}(T_i, Y_i)}}{\text{Var}(T_i)} \right)$$

where $\bar{Y}_1^{(T)} := \frac{1}{m_1} \sum_{i=1}^{m_1} Y_i$; $\bar{Y}_0^{(T)} := \frac{1}{m_0} \sum_{i=1}^{m_0} Y_i$, and m_1 and m_0 are the numbers of participants *receiving or not receiving* the intervention respectively. This is unbiased for ATE if it can be assumed that mean potential outcome is exchangeable between levels of treatment receipt. This is a very strong assumption because it is highly likely that a baseline variable (for example severity of illness) drives treatment receipt and also affects outcome. Figure 4.1 includes such an unobserved variable as variable U, which

would commonly be referred to as a confounder of the effect of T on Y. Therefore this estimator of β is likely to be subject to bias. Another way of stating this is that the estimator, which throws away the protection of random allocation, invites bias because the assumption of exchangeability is implausible (and is also impossible to test). Observed confounders could be conditioned on, making the estimator unbiased under the less restrictive assumption of conditional mean exchangeability.

The *per protocol estimator* estimates efficacy based on the result of random assignment, but only for those who adhere to their allocation (i.e. it aims to estimate ATT). This is parameter β in Figure 4.1 for the population who receive active intervention when offered it. The per protocol estimator for a continuous outcome is:

$$\widehat{\text{Per protocol}}\ \text{effect} := \bar{Y}_1^{(P)} - \bar{Y}_0^{(P)}$$

where $\bar{Y}_1^{(P)} := \frac{1}{h_1}\sum_{i=1}^{h_1} Y_i$; $\bar{Y}_0^{(P)} := \frac{1}{h_0}\sum_{i=1}^{h_0} Y_i$, and h_1 and h_0 are the subpopulations who *receive or do not receive* treatment when they have been offered/not offered the treatment respectively. The treatment effect is observed as a contrast between those who were allocated to intervention and complied versus those in the control arm who received the control treatment. This estimator is unbiased if it is assumed that mean potential outcome is exchangeable between levels of treatment allocation and treatment receipt. For similar reasons to the bias inherent in the as-treated estimator above, this estimator is also likely to suffer from bias. In summary, both of these methods are likely to suffer from selection bias because of the influence of unobserved confounding between treatment receipt and outcome, as shown in Figure 4.1 (Hernán and Hernández-Díaz, 2012; White, 2005).

The major problem which any efficacy analysis must address is the presence of unobserved confounding (non mean exchangeability) between treatment receipt and outcome. The as-treated and per-protocol effects do not achieve this. However, there are some methods that can do this and these will be introduced in Section 4.4.

4.4 Randomisation-based efficacy estimators for addressing non-compliance

This section will review existing randomisation-based efficacy estimators that can be used in trials with non-compliance, i.e. non-receipt of the treatment within the active arm.

This section of the chapter explores these estimators in the simple scenario of binary treatment receipt and no contamination, for example when the active treatment is not available other than when offered. I will focus on two major modelling frameworks for addressing non-compliance: principal stratification and instrumental variables methods.

4.4.1 Principal stratification

Principal stratification is a general framework that is most commonly applied in experimental settings. The method enables the estimation of principal effects (effects of random treatment allocation) within strata, which are defined by the levels of binary treatment allocation and binary treatment receipt (Frangakis and Rubin, 2002), using structural equation modelling. These strata are partially latent because complier status cannot be observed for a given participant. The number of principal strata tends to be four, the product of two levels of allocation and two possible levels for treatment receipt, as shown earlier in Table 4.2. Principal stratification can provide an estimator of CACE for binary treatment receipt under various assumptions. In theory it would be possible to define more than four strata but this would lead to a need for additional assumptions in order for principal effects to be identified.

Identifiability assumptions

In reality the observed compliance statuses, as shown in Table 4.3, are very different to the latent compliance classes. The observed compliance statuses (including contaminators for the moment) do not uniquely map onto the latent compliance classes. In fact each observed compliance status comprises two latent classes.

Table 4.3: Observed compliance status.

R	T	**Compliance status**	**Possible latent compliance class membership**
0	1	Control contaminator	Always taker or defier
0	0	Control non-contaminator	Complier or never taker
1	1	Treatment complier	Complier or always taker
1	0	Treatment non-complier	Never taker or defier

Assumption A-S1: Latent class membership is exchangeable between trial arms

In order to identify CACE it is necessary to make some assumptions. Latent compliance class can be assumed to be independent of random treatment offer, or exchangeable (i.e. $\mathrm{E}\big[\{T_i(R=0), T_i(R=1)\}|R_i=1\big] = \mathrm{E}\big[\{T_i(R=0), T_i(R=1)\}|R_i=0\big]$). This is part of the assumption of full exchangeability, as described earlier, which is reasonable given the use of random assignment. This has the noteworthy impact that latent compliance class can be handled as a pre-randomisation variable. This variable can be considered as a latent effect modifier, allowing the treatment effect to vary between different levels of treatment receipt.

Assumption A-S2: Monotonicity

It is commonly assumed that there are no defiers within the population, which is known as the assumption of *monotonicity* ($T_i(1) - T_i(0) \neq -1$; Imbens and Angrist, 1994). That is, there are no people who take the opposite treatment to whatever they are offered. This simplifies matters greatly because it means that only two latent classes exist: the compliers and never takers (the lack of contaminators implies no always takers). These can be directly observed in the active intervention group (observed treatment compliers and treatment non-compliers, respectively, in Table 4.3) but not the control group.

Assumption A-S3: Exclusion restriction

Finally, it is necessary either to make an assumption about the impact of treatment among never takers or to find predictors of latent compliance status (see assumption A-S4 later). Regarding the former, it can be assumed that the effect of treatment among never takers is zero – this is the so-called *exclusion restriction*. Put another way, this assumes that the effect of allocation on outcome is entirely conferred through the receipt of this treatment, i.e. the offer of treatment has no effect if it does not affect the treatment received. CACE is consequently identified on the basis of the exclusion restriction, the assumption that the proportion of latent compliers is the same in the trial arms (assumption A-S1), and the assumption of no defiers.

Definition of stratification CACE estimator

Principal stratification provides an estimator which is constructed by fitting structural equation models using maximum likelihood. Under assumptions A-S1–A-S3 this estimator

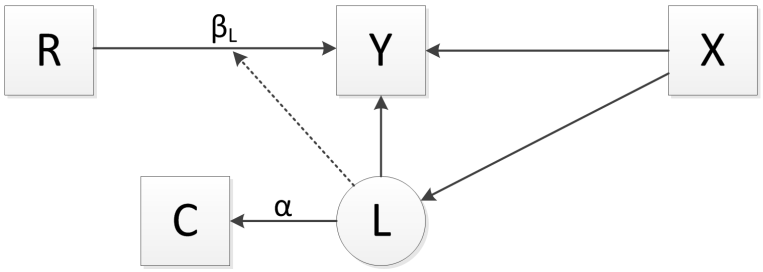


Figure 4.2: Structural equation model diagram illustrating use of principal stratification to address treatment non-compliance. Observed compliance status (C; coded one for observed compliers in the active intervention arm, zero for observed non-compliers in the same trial arm, and missing values for controls) is perfectly predicted by latent compliance status (L; i.e. $\alpha = 1$) due to the fact that latent compliers and never takers can be observed as compliers and non-compliers respectively in the active intervention arm. Latent compliance status is an effect modifier of the relationship between trial arm and outcome. Parameter β_L is constrained to zero for never takers (i.e. $\beta_n = 0$) and is freely estimated for compliers to provide an estimate of CACE (i.e. $\beta_c = CACE$).

(here named **E-STR1**) is unbiased for CACE. See Figure 4.2 for an illustration of what is being estimated with an explanation of the parametrisation of the model in the caption. CACE can be estimated using a mixture model, which allows a latent factor to predict subpopulation membership and therefore treatment effects within these strata. This provides a consistent estimator for CACE and a valid standard error for large samples. The model can be identified and parameters estimated in any package that allows structural equation models (SEMs) with latent classes (e.g. in Stata with the 'gllamm' package or in MPlus).

Allowing for baseline predictors of outcome and compliance status

Baseline covariates (X in Figure 4.2) that predict outcome are useful for increasing the efficiency of the CACE estimate for a continuous outcome (White, 2005). In other words, such prognostic covariates increase the power of the estimate. Baseline predictors of latent compliance status are useful for enhancing model identifiability and gaining precision. Such pre-randomisation variables can be found by searching for predictors of observed compliance status. When treatment is not available to those in the control arm, the search for these variables is performed within the active intervention arm. These variables are then used as covariates to predict class membership in the latent class model.

Assumption A-S4: Prediction of compliance status by baseline variables

If compliance status can be predicted this allows assumption A-S3 to be relaxed and therefore the treatment parameter for never takers to be estimated. This was demonstrated by Dunn et al. (2012) when estimating the effects of different levels of cognitive behavioural therapy (full, partial, none) compared to treatment as usual for people with psychosis. The authors tightened and then relaxed the exclusion restriction in order to check the sensitivity of the estimate for those who received full therapy. As the number of principal strata increases so does the number of predictors needed in order to distinguish between these subpopulations.

Missing outcome data

It is highly likely that a trial of an intervention in mental health research will be subject to missing outcome data and possible that the missingness will be extensive, despite the best efforts of those running the trial. The most appropriate analytical method for addressing missing data depends on the missing data generating mechanism. In the language of Little and Rubin (2014), when the mechanism is "ignorable" then the latent class model can assume either missing at random (MAR) or latent ignorability (LI). Under MAR, it is assumed that the probability of missingness is independent of outcome, given observed outcome data ($\Pr(N|Y_o, Y_m) = \Pr(N|Y_o)$, where the indices m and o represent missing and non-missing outcome). The CACE estimator **E-STR1** is valid under MAR in Figure 4.2. Under the weaker assumption of LI, the probability of an outcome value being missing is considered to be independent of outcome, given observed outcome and latent compliance status ($\Pr(N|Y_o, Y_m, L) = \Pr(N|Y_o, L)$; Rubin and Little, 2002). The stratification estimator has been shown to be valid under LI in Figure 4.3. These assumptions can be relaxed by covarying on any pre-randomisation variable that can predict missingness (Pickles and Croudace, 2010).

Assumption S5: Compound exclusion restriction

In order for the model to be identified under LI an additional ("compound") exclusion restriction must be specified. This states that for those participants whose treatment is different to random allocation, the offer of treatment has no direct impact on missingness. This is illustrated in Figure 4.3.

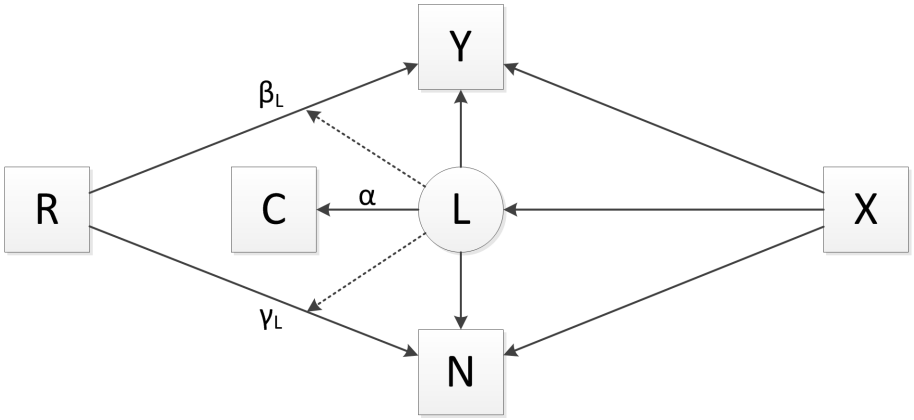


Figure 4.3: Structural equation model diagram illustrating principal stratification with missing outcome values assumed to be ignorable conditional on latent compliance status. This model is similar to Figure 4.2 but now allows the effect of treatment on response (N) to be modified by latent compliance status. This parameter (γ_L) is assumed to be zero for never takers and is freely estimated for latent compliers.

Summary

Principal stratification allows the estimation of the causal effect of latent treatment receipt on outcome. It enables assumptions regarding missing value generating processes to be relaxed but is restricted to situations where there are finite numbers of latent classes. Extensions of this framework, for example to more than four principal strata due to further levels of latent compliance status, demand predictive baseline covariates in order to allow identifiability (Emsley and Dunn, 2012). The next section will show how another method, estimation using IVs, can be used to extend causal estimation to other models of the relationship between treatment and outcome.

4.4.2 Instrumental variables

In a linear model all covariates can be classified as either *endogenous* or *exogenous*, terminology that arose in the economics literature (Wooldridge, 2010). An endogenous covariate is a predictor variable that is correlated with the error term of the outcome. A predictor variable that is unrelated to the error term is known as an exogenous variable. Note that the classification of whether a covariate is endogenous or exogenous is dependent on how the model is specified. If all the covariates are exogenous then the ordinary least squares (OLS) estimator of all covariates on outcome estimates the causal effect of each of these. If a covariate is endogenous then the OLS estimate of its effect is biased.

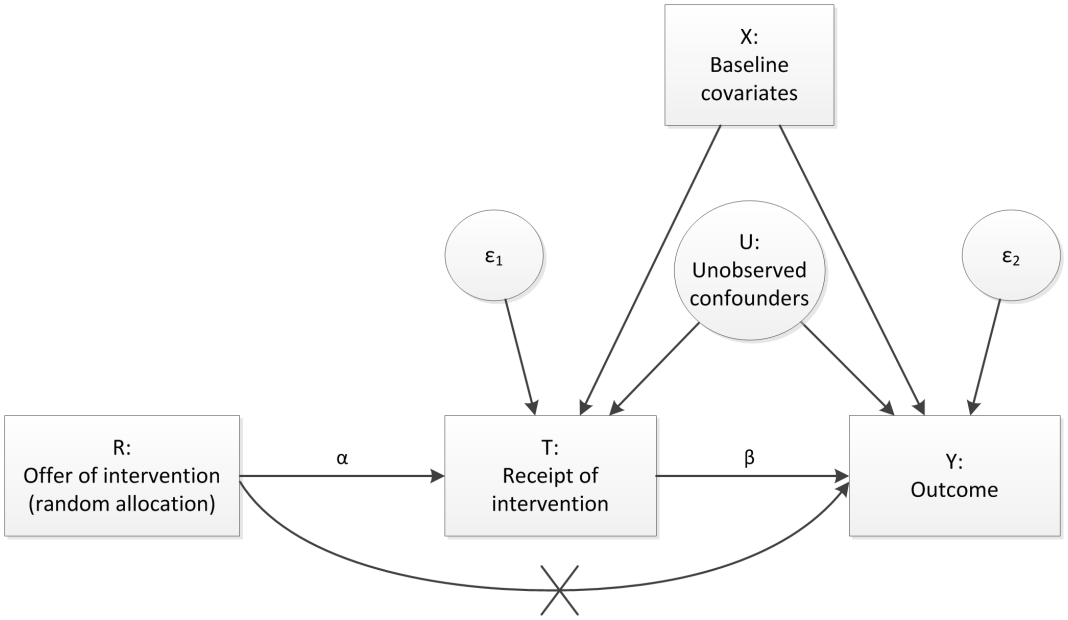


Figure 4.4: Structural equation model diagram assumed to estimate the causal effect of treatment receipt on outcome (parameter β).

The method of instrumental variables, which was developed in the econometrics literature, provides an approach for estimating the causal effect of one variable on another that accounts for confounding between these variables. The effect of such confounding is to induce a correlation between the predictor variable and error term of the outcome (i.e. the predictor is endogenous). A variable that is correlated with the endogenous variable and not with the error term of the outcome (i.e. exogenous) provides what is known as an *instrument*. In the context of treatment compliance within a randomised controlled trial, the endogenous variable is treatment receipt and the parameter of interest is its causal effect on patient outcome, as shown in Figure 4.4.

Formally, the definition of an IV includes three parts, which are known as the core conditions. As will be seen, random treatment allocation can be considered to be a suitable candidate for the choice of instrumental variable for the effect of treatment receipt in the context of a randomised controlled trial.

Assumption A-I1: Inclusion restriction

One of these core conditions, which is often called *relevance* or the *inclusion restriction*, is that the set of IVs must be predictive of the vector of endogenous variables, such that the predictions are not linearly related. In order to achieve this, at least as many IVs as endogenous variables are needed. In the current application there is one endogenous

variable, T, and one IV, R, and therefore the assumption is $R \not\perp T$. In Figure 4.4 the parameter α represents the predictive effect of R on T. This core condition is the only one which is testable.

Assumption A-I2: Independence of R and error term of Y

It is assumed that the set of instruments are statistically independent of the error term of the linear model. In the model represented by Figure 4.4 this means that R must be independent of ϵ_2. It is assumed that there is no direct pathway from the set of instruments to outcome and no common cause of both. This is fulfilled by the following two assumptions.

Assumption A-I2a: Exclusion restriction The assumed absence of a direct relationship between instruments and outcome is known as the exclusion restriction. This is equivalent to assuming that parameter $\beta = 0$ for never takers in the figure.

Assumption A-I2b: No common cause of instrument and error term The final core condition is that the set of IVs needs to be statistically independent of latent confounders between the endogenous variables and outcome. In Figure 4.4 this means that the instrument and outcome must not share any common causes ($R \perp U$). This is fulfilled in a trial because randomisation ensures no drivers of R.

In a trial, assumptions A-I1 and A-I2b are met due to randomisation but the appropriateness of A-I2a needs to be considered. In a trial where either participants or clinicians are not blind it is possible that the expectations of participants could lead to the exclusion restriction being violated. For example, the disappointment of participants who are allocated to the control arm may lead to discouragement or even feelings of retaliation. This phenomenon is thought to lead to poorer patient outcomes and is known as resentful demoralisation (Onghena, 2005). Another possible weakness in the assumptions of IV methods arises when treatment receipt is defined as binary when in fact it is continuous. In this case non-compliers who receive a small dose are incorrectly assumed to have received nothing.

Assumption A-I3: Treatment heterogeneity

It is necessary to make a further assumption to allow identification of the causal effect of T on Y. It can either be assumed that the effect of treatment is constant or that

there are no defiers in the population. There are different levels of effect homogeneity, the strongest of which is that the effect of treatment receipt on outcome is the same for all individuals (Hernán and Robins, 2018). Under this assumption, ATE is equal to CACE. A weaker effect homogeneity assumption is that the causal effect of T on Y is the same between levels of treatment offer in observed compliers and non-compliers ($\mathrm{E}\big[Y_i(T=1)-Y_i(T=0)|R_i=1,T_i=t\big]=\mathrm{E}\big[Y_i(T=1)-Y_i(T=0)|R_i=0,T_i=t\big]$ for t=0,1). Put another way, this means that the average causal effect of T on Y is equal within levels of treatment receipt. The major weakness in the plausibility of this assumption, as pointed out by Hernán and Robins (2018), is the possibility of effect modification of the causal effect by unmeasured confounders. An alternative assumption is to assume that there are no individuals within the population who would always receive the opposite treatment to what they are offered (no defiers; Imbens and Angrist, 1994). This assumption of monotonicity (described earlier) combined with the assumption of no treatment effect within the never takers (assumption A-I2a) implies that the the IV estimand represents the effect of treatment within latent compliers (i.e. CACE, as defined in Equation 4.1). In this manner the IV estimand is a type of LATE.

Definition of IV estimator

In order to estimate CACE, as defined earlier, an instrumental variables approach assumes that there are two levels of latent treatment receipt, latent compliers ($T_i(1)-T_i(0)=1$) and everyone else ($T_i(1)-T_i(0)\neq 1$). Under assumption A-I3 it is assumed that there are no defiers, meaning that the latent compliers must be observed compliers in the active intervention arm. When treatment receipt is binary and no controls receive treatment (no contamination), the never takers are observed non-compliers in the active intervention arm. It is then possible to estimate treatment efficacy in the following manner. First, it is assumed that potential outcome among those in the active intervention arm is a combination of outcome for latent compliers and never takers. Under assumptions A-I1 and A-I2 randomisation, R, provides a suitable instrument, Z.

$$\mathrm{E}[Y_i(Z=1)]=\mathrm{E}[T_i(Z=1)]\mathrm{E}[Y_i(Z=1,L=c)]+(1-\mathrm{E}[T_i(Z=1)])\mathrm{E}[Y_i(Z=1,L=n)]$$

where $\mathrm{E}[Y_i(Z=1,L=c)]$ is expected potential outcome under allocation to treatment and receipt of it (i.e. amongst latent compliers), and $E[Y_i(Z=1,L=n)]$ is expected

potential outcome under allocation to treatment and non-receipt of it (i.e. amongst never takers). Based on the assumption of consistency and assuming that the relationship between treatment assignment and treatment receipt is linear, this can be expressed in the active intervention arm in the observable world as:

$$\mathrm{E}[Y_i|Z_i=1] = \alpha_c \mathrm{E}[Y_i|Z_i=1,T_i=1] + (1-\alpha_c)\mathrm{E}[Y_i|Z_i=1,T_i=0] \tag{4.2}$$

where α_c is the proportion of participants who comply in the active intervention arm. The relationship between potential outcome and treatment receipt can also be expressed in the control arm:

$$\begin{aligned}\mathrm{E}[Y_i(Z=0)] = \mathrm{E}[T_i(Z=0)]\mathrm{E}[Y_i(Z=0,L=c)] + \\ (1-\mathrm{E}[T_i(Z=0)])\mathrm{E}[Y_i(Z=0,L=n)]\end{aligned}$$

However, latent compliance cannot be observed in the control arm, which therefore means that an extra assumption is required in order to express this, eventually, in observable terms. Specifically, it is assumed that the proportion of control participants who would comply had they been offered active intervention is the same as that observed in the active intervention arm. This assumption is incorporated in full exchangeability (described earlier) and is plausible given the use of random treatment allocation. Therefore, on the basis that the parameter α_c is estimable, observed outcome amongst controls is a weighted average of expected potential outcomes:

$$\mathrm{E}[Y_i|Z_i=0] = \alpha_c \mathrm{E}[Y_i(Z=0,L=c)] + (1-\alpha_c)\mathrm{E}[Y_i(Z=0,L=n)] \tag{4.3}$$

Treatment efficacy, the difference in expected outcome between trial arms for those who would comply, can be expressed as a rearrangement of Equations 4.2 and 4.3. For a continuous outcome, efficacy (β) can be estimated with an unbiased sample as follows:

$$\begin{aligned}\beta &= \mathrm{E}(Y_i(Z=1,L=c) - \mathrm{E}(Y_i(Z=0,L=c))) \\ \hat{\beta} &= \frac{((\bar{Y}_{1.}) - (1-\hat{\alpha}_c)(\bar{Y}_{10})) - ((\bar{Y}_{0.}) - (1-\hat{\alpha}_c)(\bar{Y}(Z=0,L=n)))}{\hat{\alpha}_c}\end{aligned} \tag{4.4}$$

where $\bar{Y}_{zt}$ is mean outcome given that $Z=z$ and $T=t$. This equation is not estimable

because the right side includes a potential mean outcome, $(\bar{Y}(Z=0,L=n))$, which cannot be estimated in the observable world based on assumptions made so far. However, the exclusion restriction assumes that the treatment effect amongst those who do not comply is zero, implying that mean outcome for non-compliers in the active intervention and control trial arms are exchangeable. Substituting $(\bar{Y}_{10})$ for $(\bar{Y}(Z=0,L=n))$, and simplifying by expressing $(\bar{Y}_{1.})-(1-\hat{\alpha}_c)(\bar{Y}_{10})$ as $\hat{\alpha}_c(\bar{Y}_{11})$, Equation 4.4 can be expressed in observable terms as an estimator for β:

$$\hat{\beta} = \frac{\hat{\alpha}_c(\bar{Y}_{11}) + (1-\hat{\alpha}_c)(\bar{Y}_{10}) - (\bar{Y}_{0.})}{\hat{\alpha}_c} \tag{4.5}$$

This estimator is named **E-IV1** here. The standard error for this can be found by bootstrapping.

Alternatively, the treatment efficacy parameter can be expressed as a ratio of the paths in the SEM path diagram in Figure 4.4. This ratio is:

$$\beta = \frac{\alpha\beta}{\alpha} = \frac{E[Y_i|Z_i=1]-E[Y_i|Z_i=0]}{E[T_i|Z_i=1]-E[T_i|Z_i=0]}$$

An unbiased estimator for this (here named **E-IV2**) is:

$$\hat{\beta} = \frac{(\bar{Y}|Z_i=1)-(\bar{Y}|Z_i=0)}{\Pr(T_i|Z_i=1)-\Pr(T_i|Z_i=0)} = \frac{\widehat{\text{ITT effect of Z on Y}}}{\widehat{\text{ITT effect of Z on T}}}$$

If there are no missing outcome data, Equation 4.5 simplifies to this (Dunn et al., 2003). This estimator was introduced by Bloom (1984) and is occasionally known as either the Bloom IV estimator (e.g. Dunn et al., 2005), Wald estimator (Wald, 1940), or ratio estimator.

The most commonly used methods for estimating the causal effects of endogenous variables, which were developed in the econometrics literature, involve two stages (Wooldridge, 2010). In general, these methods require that the instruments must be capable of predicting the endogenous variables so that the predictions can be used in the estimation procedure. The first stage comprises the prediction of the endogenous variable by the instrument. In the second stage the information from this prediction is used in place of the endogenous variable in order to estimate the causal parameter of interest. Methods that can handle multiple endogenous variables are reviewed in the next section. In the case where there is one endogenous variable (treatment receipt) and

the instrument is random allocation, this simplifies to Bloom's IV estimator, as already described.

In practice, treatment efficacy is often estimated using a regression approach, of which the most common examples are the two stage least squares (2SLS) and adjusted treatment response (ATR) methods. Generally speaking, in the first stage they use ordinary least squares to estimate the effect of the instrument on the endogenous variable and in the second stage use the post-estimation information from this regression to estimate the unbiased impact of the endogenous variable on outcome. In the context of treatment non-compliance (i.e. the model shown in Figure 4.4), T is initially regressed on Z and the predicted values of T are estimated and saved. Following this, Y is regressed on the predicted value of T obtained previously (IV(2SLS) method, Wooldridge, 2010). This method is named here as **E-IV3**. The second method comprises first regressing T on Z, saving the residuals from this, and then regressing Y on T and the saved residuals (introduced in the biostatistics literature by Nagelkerke et al., 2000). This is referred to here as the IV(ATR) approach but is also sometimes referred to as the *control function* method in the econometrics literature (Wooldridge, 2015) or the *two stage residual inclusion* method in health economics (Terza et al., 2008). Provided the causal effect is linear, the two approaches are equivalent and therefore produce the same estimate of the effect of treatment receipt on outcome. IV estimation can be performed using existing packages in Stata ('ivregress') or R ('tsls' or 'ivreg'). The 2SLS procedure fits the model in one go and provides model-based standard errors. In contrast, the methods involving creation of predictions of treatment receipt (or residuals) must use bootstrapping to generate correct standard errors in the second stage.

Maracy and Dunn (2011) show that these approaches are equivalent to another two-stage method named *structural mean modelling*. All three methods are based on the same model for the relationship between receipt of treatment and treatment effect. Using this type of modelling, the treatment efficacy parameter can be estimated by g-estimation, as developed by Robins (1994) and Fischer-Lapp and Goetghebeur (1999). This review will not go into much detail about structural mean models but some basic information about the overlap between these and IV methods is given here briefly. The stages of g-estimation involve predicting potential outcomes (of the endogenous variable and outcome) under treatment and control separately for all participants and then regressing the difference between these on the predicted values of the endogenous variable. The

basic rationale behind g-estimation and IV methods is that random allocation can be used to create strata that are conditionally exchangeable, within which it is possible to estimate causal treatment effects.

There are other IV estimation approaches apart from the two-stage methods. Full information maximum likelihood (FIML), sometimes shortened to simply maximum likelihood (ML), is one such alternative (Imbens and Rubin, 1997; Jo and Muthén, 2001). Under this estimation technique, outcome is predicted by latent compliance status and an interaction between it and random treatment group allocation. The model is identified if the effect for latent non-compliers is assumed to be zero (i.e. the exclusion restriction) or if there are baseline variables that predict latent compliance status membership. The model is then fitted using the expectation maximisation algorithm (Dempster et al., 1977). Another estimation approach is limited information maximum likelihood (LIML) (Anderson and Rubin, 1949).

Multiple endogenous variables and multiple instruments

In order for a model to be identified when more than one covariate is endogenous, the number of instruments needs to be at least as large as the number of endogenous variables (Dunn and Bentall, 2007). This means it is necessary to search for potential instruments. Any candidate instrument must meet the core conditions, meaning that an instrument must be relevant to the intermediate variable, not directly related to outcome (other than through the endogenous variable) and not share common causes with outcome. In practice, this means that instruments must be pre-randomisation covariates that are predictive of the endogenous variables but not effect modifiers of the causal relationship between the endogenous variables and outcome. Possible candidates for these in trials are interaction terms between such predictive covariates and random allocation. The interaction with randomisation fulfils the assumption that the IV and outcome have no common cause. It must still be assumed that there is no direct effect of the instrument on outcome in order for the interaction term to satisfy all core conditions necessary for being an IV. This assumption can be a strong one in trials where blinding is not possible, as described earlier.

For example, Goldsmith et al. (2015) evaluated the causal effect of two endogenous variables, number of treatment sessions and the interaction between sessions and therapist alliance, in order to investigate the impact on treatment effect of sessions attended (at

particular levels of alliance) and of a one-unit increase in alliance when therapy took place. They found that duration of illness, years of education, outcome score at baseline, and treatment centre predicted number of treatment sessions attended. They then created interaction terms of these variables with random allocation and used these as IVs. The participant-level predictions of sessions and the interaction between sessions and alliance were then used to estimate the effect of the endogenous variables on outcome.

Dose-response relationships for continuous treatment receipt

Up until this point it has been assumed that participants either receive or do not receive the active intervention, and that the effect of intervention is zero amongst those who do not receive it. It is possible to extend IV methods to include a dose-response relationship between a continuous measure of dose of treatment (e.g. number of sessions attended) and treatment effect, as shown by Maracy and Dunn (2011). In this case, the model in effect stratifies the population (based on dose of treatment received if offered) and then estimates the change in treatment effect between the subpopulations.

For instance, the causal effect of treatment offer ($\Delta_{r,i}$) can be modelled with a linear dose-response relationship (Maracy and Dunn, 2011):

$$\Delta_{r,i}|(D_i(R=1)=d) = \beta d + \epsilon_i \tag{4.6}$$

This constitutes the treatment effect resulting from attending d sessions (only taken if offered) where the causal parameter β represents the change in treatment effect due to a one-unit increase in dose, i.e. $\beta = \mathrm{ACE}_{\mathrm{d}_1+1,\mathrm{d}_0} - \mathrm{ACE}_{\mathrm{d}_1,\mathrm{d}_0}$. This demonstrates that the latent complier population has been split into strata defined by the level of d_1. Never takers remain the same as for binary treatment receipt ($D_i(R=0)=0$, $D_i(R=1)=0$). The residual term ϵ represents unaccounted variability in the treatment effect in a subpopulation defined by the level of d.

Certain assumptions must be made in order for the causal effect of dose on outcome to be identified. The assumptions of the inclusion restriction (assumption A-I1), the exclusion restriction (assumption A-I2a), and no common cause of instrument and the error term of the outcome (assumption A-I2b) are made. The meaning of the exclusion restriction here is that the treatment effect is zero for those who would attend no sessions. This is represented by the lack of an intercept term in Equation 4.6 and the assumption that the expectation of the residual term is zero. The final identifiability assumption

relates to treatment heterogeneity and includes two parts. The first is a more general monotonicity assumption which now assumes that the endogenous variable is a non-decreasing function of the instrument. In the context described here, this means that the offer of treatment is assumed not to lead to any participant attending fewer sessions of the treatment compared to what the patient would have done had he or she not been offered the treatment ($D_i(R = 1) \geq D_i(R = 0)$). The second part is the assumption that the endogenous variable and residual term in Equation 4.6 are not related ($\text{Cov}(D, \epsilon) = 0$). In econometrics this is known as the assumption of no *essential heterogeneity*. This means that it must be assumed that there is no unmeasured confounding between these two terms.

This linear model can be easily extended to allow a non-linear dose-response relationship by including further terms. For example, Maracy and Dunn (2011) modelled treatment effect with linear and quadratic terms for number of treatment sessions, therefore allowing this relationship to be non-linear. This type of modelling relies on an expectation or expert theory about the relationship between the exposure and outcome. Burgess et al. (2014) introduced a non-parametric method for stratifying the population and estimating strata-specific causal effects of exposure. This method can therefore be used to investigate the shape of the relationship between exposure and outcome. The approach is to stratify the population based on residual variation in the exposure after conditioning on the instrument. The authors then used the ratio method to estimate the local average causal effect of exposure on outcome within these strata. A more general method, which the authors also describe, is to order individuals according to the values of the residuals found after regressing the exposure on the instrument and then use a sliding window to estimate the local average causal effect within these windows. Estimates can then be plotted against median exposure level (within each window) to explore the shape of the exposure-outcome relationship. The choice of window size (i.e. number of observations) influences the profile of the plot: a smaller size leads to a sharper resolution in the estimated shape of the relationship; a larger window provides greater precision. The drawbacks of these methods is that they require a large number of observations and data from a wide range of exposure values.

The IV methods described above (IV(2SLS) or IV(ATR)) can be used to explore the relationship between continuous exposure and outcome and to estimate the change in the causal efficacy parameter (i.e. $\text{ACE}_{\text{d}_1+1,\text{d}_0} - \text{ACE}_{\text{d}_1,\text{d}_0}$ where $d_0 = 0$, defined earlier).

The bottom line is that such methods can estimate the change in treatment effect for each additional dose (e.g. session) of treatment received. Where there is more than one dose variable, further instruments must be found as described in the context of multiple binary endogenous variables. The dose-response estimator is referred to here as **E-IV4**.

Allowing for predictors of outcome

Baseline covariates which predict outcome can be included in the two stage least squares methods. These covariates are included as explanatory variables in both stages of the methods. They are typically variables which are anticipated to be predictive of outcome (such as the outcome variable measured pre-randomisation) or variables which predict outcome by design (e.g. randomisation stratifiers). Commonly the effect of such covariates on outcome is constrained to be the same in the various compliance classes, especially if they are included as effect modifiers in the prediction of endogenous variables.

Missing outcome values

So far it has been assumed that there are no missing outcome data, which in effect makes estimators **E-IV1** (modified Bloom/ratio estimator with proportions of observed compliance and non-compliance predicting outcome), **E-IV2** (Bloom/ratio estimator), **E-IV3** (2SLS estimator for binary observed compliance), and **E-IV4** (2SLS estimator for continuous treatment receipt in intervention arm) equivalent. When there are missing outcome data, it is necessary to make assumptions regarding the missing data generating mechanism under which the estimator remains valid. Of the four estimators described in this section, **E-IV2**, **E-IV3** and **E-IV4** make the strongest assumptions regarding the mechanism leading to missing data. They assume that, given observed outcome data, the missingness generating mechanism does not depend on the missing data ($\Pr(N|Y_o, Y_m) = \Pr(N|Y_o)$). This is to assume that missingness is ignorable and that missing data are MAR. Estimator **E-IV1** additionally allows observed compliance to predict missingness ($\Pr(N|Y_o, Y_m, C) = \Pr(N|Y_o, C)$).

Another method for addressing missing data that accounts for the process that generated the incomplete data is MI (originally proposed by Rubin, 1978, 2004). In general, MI is a technique for replacing missing data with a range of plausible values; this provides a series of imputed datasets whose analyses are pooled together to provide estimates

that are valid under the expected missing data generating mechanism. The method includes two main stages: imputation and analysis. In the imputation step missing data are replaced by values that maintain associations between variables in the model whilst at the same time reflecting the uncertainty of these predictions. This can be done using chained equations (a parametric approach introduced by Van Buuren et al., 1999) or Monte Carlo Markov chains which assume multivariate normality (Schafer, 1997). In the second step, the series of imputed datasets are each analysed and the results are combined using Rubin's rules (Rubin, 2004). In practice, MI can be used to allow predictors of missingness (including observed compliance) and therefore relax the assumptions of estimators **E-IV2**, **E-IV3**, and **E-IV4**.

Accounting for clustering

It is possible that outcome data could demonstrate lack of independence between individuals' observations, for example due to groups of participants sharing therapists. The prognosis of participants who share a therapist might be related due to shared treatment techniques. An estimation model for outcome that does not address this assumes that data are independent and therefore will (usually) tend to underestimate standard errors. In this section I review existing methods that account for clustered outcome data and specify how I will apply these to the ITT and IV estimators that I have described.

It is possible to take an estimation approach that assumes the existence of a multi-level structure within the data, i.e. levels of residual error terms for participants and therapists. Mixed-effects models can be used to handle such data dependency and therefore provide correct estimation of standard errors when all covariates can be considered exogenous. These models allow the hierarchical data structure to be acknowledged within the procedure for estimation of coefficients by modelling error terms at all levels. Let us return to the **E-ITT** estimator and assume a linear model for outcome Y with random treatment allocation R, outcome measured at baseline X, level-two residual error term ω_j (with subscript j labelling clusters and i labelling patients within the j^{th} cluster), and level-one residual error term ε_{ij}:

$$Y_{ij} = \alpha + \beta_1 R_{ij} + \beta_2 X_{ij} + \varepsilon_{ij} + \omega_j$$

In this model, terms ε_{ij} and ω_j can be modelled as random effects, i.e. terms that take expectation zero and some variance. Coefficients for intercept α, slopes β_1 and

β_2, variances and covariances can be estimated using an estimation procedure such as maximum likelihood.

Generalised two stage least squares (G2SLS) method

A collection of methods exist that enable estimation of parameters where data take this multi-level structure and where some covariates are endogenous. These are generalisations of the 2SLS estimator. One of these, the G2SLS method, passes exogenous variables through the feasible generalised least squares transformation and then uses the results of this as instruments (Balestra and Varadharajan-Krishnakumar, 1987). Endogenous variables and outcome are also passed through the same transformation, enabling unbiased and consistent estimation of coefficients, variances and covariances using the familiar 2SLS approach. Error terms are handled as random variables that are assumed to be independent and identically distributed over clusters. The G2SLS method can be performed in Stata using the 'xtivregress' command and specifying random error ('re') in the options. In R it can be done using the 'plm' function with the Balestra and Varadharajan-Krishnakumar ('bvk') method. I will use the G2SLS approach as a method of accounting for clustering in the efficacy analysis of simulated trials with measures of treatment receipt in Chapter 5.

Clustered robust variance estimator approach

A separate approach is to use OLS estimation and account for lack of independence between outcome data by allowing for correlations within groups. Observations are then assumed to be independent between groups (but not within them), i.e. residual errors are correlated within clusters but not across them. This method relies on the use of robust standard errors (White, 1984). I will briefly describe how the robust variance estimator is derived. I return to the OLS estimator for outcome Y, with independent variables random treatment allocation R and outcome measured at baseline X. For the moment I assume there are participant-level residual errors (ε_i) and no level-two errors. I refer to the vector of the dependent variable as $\mathbf{y}$ and the $2 \times n$ matrix for the independent variables as $\mathbf{X}$. The estimator for the effect of random treatment allocation on outcome (β_1) is,

$$\hat{\beta_1} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

And the robust variance estimator for this coefficient is,

$$\begin{aligned}\text{var}_{\text{rob.}}(\hat{\beta}_1) &= (\mathbf{X}'\mathbf{X})^{-1}\text{var}(\mathbf{X}'\mathbf{y})(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\Big(\sum_{i}^{n}\varepsilon_i^2\mathbf{x}_i'\mathbf{x}_i\Big)(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

I now return to the linear model for outcome with level-one residual error ε_{ij} and level-two residual error ω_j. The clustered robust variance estimator for the coefficient is similar to the robust variance estimator but with residuals replaced by their sums over each cluster,

$$\begin{aligned}\text{var}_{\text{clust.}}(\hat{\beta}_1) &= (\mathbf{X}'\mathbf{X})^{-1}\text{var}(\mathbf{X}'\mathbf{y})(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\Big(\sum_{j}^{n_j}u_j'u_j\Big)(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

where n_j is the number of clusters and $u_j = \sum_j^{n_j}\varepsilon_i\mathbf{x}_i$. I will use the clustered robust variance estimator in conjunction with IV estimators in the efficacy analysis of D6 in Chapter 7.

Summary

The IV framework provides a flexible parametric approach for modelling treatment efficacy. This section described how, in the context of non-compliance (treatment non-receipt in the active intervention arm), IVs can be used to estimate the causal effect of treatment. This included treatment on dichotomous and continuous scales, and circumstances in which there could be more than one measure of treatment receipt (multiple endogenous variables). The use of IVs to identify causal parameters relies upon a large number of assumptions, some of which are untestable. These include the three core conditions of IVs (relevance, exclusion restriction, and no common cause of instrument and error term of outcome), an assumed parametric relationship between the endogenous variable and outcome (often assumed to be linear), the same proportion of compliers in the active intervention and control arms, and, for a dose-response relationship, no treatment effect heterogeneity.

In comparison to principal stratification, IV methods enable more flexible modelling of causal parameters because they are less restricted by the definitions of strata. However, they make stricter assumptions about missing data because any predictors of missingness

that are included in the model must be observable variables.

4.5 Randomisation-based efficacy estimation approaches for addressing contamination

So far the efficacy analysis approaches have assumed that there is no contamination, implying the existence of only three latent classes (compliers, never takers, defiers). I will now expand the methodology and assumptions to accommodate contamination (i.e. the presence of always takers). My novel contribution will be an analysis method for estimating efficacy with continuous measures of treatment receipt in both active and control arms, i.e. in the presence of both non-compliance and contamination.

4.5.1 Principal stratification

As described in Section 4.4.1, principal stratification provides a framework for the identification of causal effects within population strata, which are latent and defined by levels of treatment allocation and treatment receipt. These so-called principal effects can be estimated using structural equation modelling. This provides an estimator of CACE for binary treatment receipt.

It is assumed that latent class membership is exchangeable between trial arms and that there are no defiers among the population (assumptions A-S1 and A-S2 described earlier). Where there is some receipt of the active intervention among those in the control arm and full compliance within the intervention arm, it is also assumed that the effect of treatment is zero amongst the stratum of participants who would always receive treatment irrespective of what they are offered. This is the exclusion restriction (assumption A-S3) and now applies to always takers rather than never takers (as was specified earlier). Where there is contamination and non-compliance, it is assumed that the treatment effect is zero within always takers and never takers.

Definition of stratification CACE estimator

Principal stratification allows CACE to be estimated by fitting a structural equation model in the context of a mixture model, where parameters and inference can be calculated using maximum likelihood. Under assumptions A-S1, A-S2, and the revised A-S3 assumption, this estimator (named **E-STR2**) is unbiased for CACE. Figure 4.5 illustrates how the

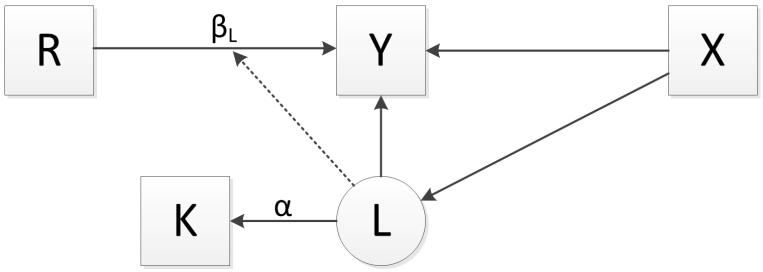


Figure 4.5: Structural equation model diagram illustrating the use of principal stratification to address treatment contamination. Observed contamination status (K; coded one for observed contaminators in the control arm, zero for observed non-contaminators in the same trial arm, and missing values for those in the active intervention arm) is perfectly predicted by latent compliance status (L; i.e. $\alpha = 1$) due to the fact that latent compliers and always takers can be observed as control non-contaminators and control contaminators. Latent complier status is an effect modifier of the relationship between trial arm and outcome. Parameter β_L is constrained to zero for always takers (i.e. $\beta_a = 0$) and is freely estimated for compliers to provide an estimate of CACE (i.e. $\beta_c = CACE$).

model is specified. With a representative sample, CACE can be estimated on the basis that latent compliance status, which perfectly predicts observed contamination status, is an effect modifier of the effect of treatment allocation on outcome. As before, this provides a consistent estimator and valid standard error for the parameter.

Where there is both treatment contamination and non-compliance, a similar approach can be taken in order to estimate CACE. Under the assumptions A-S1 and A-S2, and now assuming no effect of treatment in both always takers and never takers, the SEM mixture model provides an unbiased estimator (named **E-STR3**) for CACE. Figure 4.6 shows how observed contamination and compliance status are predicted by latent compliance status, which interacts with random treatment allocation to provide an estimate for the effect of this among latent compliers.

Allowing for baseline predictors of outcome and compliance status

Predictors of outcome, contamination status, or both can be incorporated into the model. Prognostic covariates increase the precision of the CACE estimate, whilst predictors of contamination aid model identifiability. When searching for predictors of contamination, K should be coded in a similar manner to manifest C above (in particular with missing values for those in the active intervention arm). If contamination can be predicted by the vector of baseline covariates (revised assumption A-S4), this enables the exclusion restriction to be relaxed.

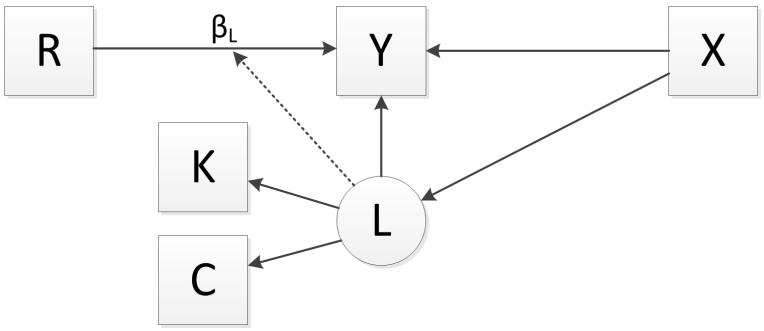


Figure 4.6: Structural equation model diagram illustrating the use of principal stratification to address treatment contamination and non-compliance. Observed contamination status (K; 1=observed contaminator, 0=observed non-contaminator, missing for those in the active intervention arm) and observed compliance (C; 1=observed complier, 0=observed non-complier, missing for those in control arm) are perfectly predicted by latent compliance status (L). Always takers and never takers can be observed as control contaminators and treatment non-compliers respectively; the residue are latent compliers. Latent compliance status is an effect modifier of the relationship between trial arm and outcome. Parameter β_L is constrained to zero for always takers and never takers (i.e. $\beta_{a,n} = 0$), and is freely estimated for compliers to provide an estimate of CACE (i.e. $\beta_c = CACE$).

Where there is both treatment contamination and non-compliance, the vector of baseline covariates must be capable of predicting always takers and never takers. For example, Hirano et al. (2000) performed a secondary analysis of a randomised controlled trial of the encouragement of clinicians to provide influenza vaccine to adults at high risk of the disease. In such an unblinded trial it is known that there will be substantial subpopulations of clinicians who always and never give the treatment to those at high risk. The researchers found that two covariates predicted both always takers and never takers. The authors used this to relax the exclusion restrictions amongst always and never takers and concluded that there was evidence for the direct effect of random treatment allocation on outcome. The sets of covariates that distinguish always and never takers from latent compliers can be the same, as in this example, or can be different.

Missing outcome data

The models in Figures 4.5 and 4.6 make the assumption that any missing data are MAR. This is because they assume that missingness is predicted by treatment allocation and any baseline covariates that are included. The weaker assumption of LI allows latent compliance status to predict missingness. In the context of treatment contamination, a further assumption is necessary in order for the model to be identified under LI. The compound exclusion restriction (revised assumption A-S5) assumes that the effect of

random treatment allocation on missingness is zero amongst always takers, allowing the parameter to be freely estimated for latent compliers. Where there is both treatment contamination and non-compliance, this parameter is assumed to be zero for both always takers and never takers.

Summary

Principal stratification provides a framework for estimation of the causal effect of latent treatment receipt on outcome and a flexible method for addressing missing data. The framework enables a stratification estimator to calculate treatment effect amongst latent compliers in the analysis of a trial with non-compliance and contamination. Baseline predictors of non-receipt of treatment in the intervention arm and receipt of treatment in the control arm provide a gain in the precision of the estimated causal treatment effect. The next section will demonstrate and extend the use of instrumental variables methods to calculate causal treatment effect in other models of the relationship between exposure and outcome.

4.5.2 Instrumental variables

Treatment contamination is a process variable that occurs after randomisation. This means that randomisation provides no protection against the possibility of confounding between it and outcome. More formally, treatment receipt amongst participants in the control group is likely to be an endogenous explanatory variable because it is expected that it is correlated with the error term of Y (due to the likelihood that the two variables share common causes). IV methods provide estimators for the effect of receipt of treatment on outcome whilst accounting for confounding.

Assumptions

The assumptions necessary for estimation are similar to the core conditions (assumptions A-I1 and A-I2) and assumed monotonicity (assumption A-I3) described in Section 4.4.2. The difference is that under the problem of contamination there are always takers but no never takers among the population. Where there is contamination and non-compliance then always takers can be identified in the control arm (described here as control contaminators) and never takers can be identified in the active intervention arm (treatment non-compliers). Latent compliers cannot be identified but their proportion

can be – it is simply the residual part of the population after summing together the proportions of always takers and never takers.

Estimation approaches

In order to estimate CACE in the context of treatment receipt in the control arm of a trial, the IV approach assumes two levels of latent treatment receipt, compliers ($T_i(1)-T_i(0)=1$), and everyone else, ($T_i(1)-T_i(0)\neq 1$). If full compliance is observed within the intervention arm then it can be assumed that there are no never takers. This implies that that the population is comprised solely of latent compliers ($T_i(1)-T_i(0)=1-0=1$) and always takers ($T_i(1)-T_i(0)=1-1=0$). The latent compliers and always takers can be observed as compliers and contaminators in the control arm, making it possible to estimate their proportions in the population.

Potential outcome under the offer of treatment is a weighted average of potential outcome for non-contaminators (latent compliers; L=c) and for contaminators (always takers; L=a). Under the assumption of consistency, this can be partially identified under offer of intervention in the observable world as:

$$\mathrm{E}[Y_i|Z_i=1]=\alpha_k\mathrm{E}[Y_i(Z=1,L=c)]+(1-\alpha_k)\mathrm{E}[Y_i(Z=1,L=a)] \tag{4.7}$$

where α_k is the proportion of participants who are not contaminated in the control arm (i.e. proportion of latent compliers). Similarly, potential outcome under the offer of control treatment can be identified in the observable world as:

$$\mathrm{E}[Y_i|Z_i=0]=\alpha_k\mathrm{E}[Y_i|Z_i=0,T_i=0]+(1-\alpha_k)\mathrm{E}[Y_i|Z_i=0,T_i=1] \tag{4.8}$$

Rearranging Equations 4.7 and 4.8, treatment efficacy (with a continuous outcome) can be expressed using the following estimator for an unbiased sample:

$$\begin{aligned}\beta &= \mathrm{E}(Y_i(Z=1,L=c)-\mathrm{E}(Y_i(Z=0,T=0)))\\ \hat{\beta} &= \frac{((\bar{Y}_{1.})-(1-\hat{\alpha}_k)(\bar{Y}(Z=1,L=a)))-((\bar{Y}_{0.})-(1-\hat{\alpha}_k)(\bar{Y}_{01}))}{\hat{\alpha}_k}\end{aligned} \tag{4.9}$$

where $\bar{Y}_{zt}$ is mean outcome given that $Z=z$ and $T=t$. This is not estimable because one term ($\bar{Y}(Z=1,L=a)$) cannot be estimated based on assumptions made so far.

However, under the exclusion restriction it can be assumed that mean outcome for always takers in both trial arms is exchangeable. Replacing $(\bar{Y}(Z=1,L=a))$ with $(\bar{Y}_{01})$, and simplifying by expressing $(\bar{Y}_{0.})-(1-\hat{\alpha}_k)(\bar{Y}_{01})$ as $\hat{\alpha}_k(\bar{Y}_{00})$, Equation 4.9 can be expressed in observable terms as an estimator for β:

$$\hat{\beta} = \frac{\bar{Y}_{1.} - \hat{\alpha}_k(\bar{Y}_{00}) - (1-\hat{\alpha}_k)(\bar{Y}_{01})}{\hat{\alpha}_k} \tag{4.10}$$

Similarly to the steps described in Section 4.4.2, when there are no missing outcome data Equation 4.10 simplifies to the ratio between the ITT effect of Z on Y and that of Z on T (estimator **E-IV2** from earlier).

If there is contamination and also treatment non-receipt in the intervention arm then this implies that the population includes (latent) compliers, always takers, and never takers $(T_i(1)-T_i(0)=0-0=0)$. The always takers are observed as contaminators in the control arm and never takers are non-compliers in the intervention arm. The proportion of compliers is the residue after the always and never takers have been identified. Following the assumptions and argument described above, an unbiased estimator for efficacy (named **E-IV5** here), as given by Dunn et al. (2005) is:

$$\hat{\beta} = \frac{\hat{\alpha}_{11}\bar{Y}_{11} + \hat{\alpha}_{10}\bar{Y}_{10} - \hat{\alpha}_{01}\bar{Y}_{01} - \hat{\alpha}_{00}\bar{Y}_{00}}{1-\hat{\alpha}_{10}-\hat{\alpha}_{01}} \tag{4.11}$$

where $\hat{\alpha}_{zt}$ is the estimated proportion receiving treatment t conditional on being randomly allocated to receive treatment z and $\bar{Y}_{zt}$ is mean outcome given that $Z=z$ and $T=t$. Cuzick et al. (1997) followed a similar approach to that above in order to provide an estimator for evaluating the effect of treatment receipt on a binary outcome (motivated in this case by the effect of breast cancer screening on mortality) when addressing non-compliance and contamination. The estimator they developed was for a risk ratio.

The standard error for the estimator **E-IV5** can be calculated by bootstrapping. The causal parameter and its standard error can also be estimated in a regression framework using the Bloom IV estimator (here referred to as estimator **E-IV6**), or the IV(2SLS) method as described in Section 4.4.2 (estimator **E-IV7**).

Dose-response relationships for continuous treatment receipt

Where there is continuous treatment receipt in the control arm (and a measure of it) and no observed treatment non-compliance, it is possible to estimate the causal effect of

treatment on outcome. First, it is necessary to state the causal assumptions needed for the specification of the causal parameter. These are:

C1. Monotonicity ($D_i(R=1) \geq D_i(R=0)$), i.e. there is nobody who would receive a larger dose if allocated to control compared to what they would receive if allocated to treatment;

C2. Linear dose-response model:

(a) $Y_i(R=0) = Y_b + \gamma_D D_i(R=0) + \tau_i$, where $Y_b = \mu_i + e_i$ (μ_i is baseline outcome and may be a function of some baseline covariates), τ_i is the error term, and $D_i(R=0)$ ranges from 0 to a maximum dose of g;

(b) $Y_i(R=1) = Y_b + \lambda_D D_i(R=1) + \epsilon_i$, where $Y_b = \mu_i + e_i$ (μ_i is baseline outcome and may be a function of some baseline covariates), ϵ_i is the error term, and $D_i(R=1)$ ranges from 0 to to a maximum dose of g;

C3. No effect of random treatment allocation on outcome other than through the exposure (the exclusion restriction):

$$\mathrm{E}\big\{\epsilon_i - \tau_i | D_i(R=1) - D_i(R=0) = 0\big\} = 0$$

C4. No unaccounted variability in ATEs within each level (subpopulation) of latent compliers:

$$\mathrm{E}\big\{\epsilon_i - \tau_i | D_i(R=1) - D_i(R=0) \neq 0\big\} = 0$$

C5. Exchangeability of potential dose and of potential patient outcome between levels of random treatment allocation:

(a) $D_i(R=1), D_i(R=0) \perp R_i$

(b) $Y_i(R=1), Y_i(R=0) \perp R_i$

Under assumptions C1, C2, C3, C4 and C5b, the average causal effect for a participant who would receive the full dose if offered it and would receive dose d_0 when allocated to the control condition (where the difference in potential doses is positive) is:

$$\begin{aligned}
\mathrm{ACE}_{\mathrm{d}_0} &= \mathrm{E}\big\{Y(R=1) - Y(R=0) | D(R=0) = d_0, D(R=1) = g\big\} \\
&= \mathrm{E}\big\{\lambda_D g - \gamma_D d_0 | D(R=1) = g, D(R=0) = d_0\big\} \\
&= \lambda_D g - \gamma_D d_0
\end{aligned}$$

The term $\lambda_D g$ is a constant that represents the maximum possible treatment effect when control participants do not receive any treatment ($d_0 = 0$). The causal parameter γd_0 represents the change in treatment effect due to a one-unit increase in dose under offer of control. This parameter is preceded by a minus sign which implies that the impact of receiving greater doses of treatment under offer of control is to reduce the treatment effect. Put another way, the more an individual is contaminated the less the benefit (or harm) of treatment would have been.

Continuous dose of treatment in both control and active intervention arms

Where there is a continuous dose of treatment in both the control and active intervention arms, the average causal effect if a participant receives a particular dose of treatment can be defined using assumptions C1, C2, C3, C4 and C5b:

$$\begin{aligned}
\text{ACE}_{\text{d}_1,\text{d}_0} &= \text{E}\big\{Y(R=1) - Y(R=0) | d_0 \geq 0, d_1 > d_0\big\} \\
&= \text{E}\big\{\lambda_D d_1 - \gamma_D d_0 | d_0 \geq 0, d_1 > d_0\big\} \\
&= \lambda_D d_1 - \gamma_D d_0
\end{aligned}$$

where $d_0, d_1 = \{0, 1, ..., g\}$ with $d_1 \geq d_0$. The motivation for defining the estimand in this way is to divide the latent compliers into subpopulations that are defined by what dose of treatment a participant would receive under intervention and control. This implies that the change in the causal parameters for a one-unit increase in dose when offered treatment is $\text{ACE}_{\text{d}_1+1,\text{d}_0} - \text{ACE}_{\text{d}_1,\text{d}_0} = (\lambda_D(d_1+1) - \gamma_D d_0) - (\lambda_D d_1 - \gamma_D d_0) = \lambda_D$. In Table 4.4, which provides the causal parameters at different levels of d_1 and d_0, this is equivalent to moving between rows (i.e. d_1 to $d_1 + 1$) at a particular value of d_0. The change in the causal parameter for a one-unit increase in dose under control is $\text{ACE}_{\text{d}_1,\text{d}_0+1} - \text{ACE}_{\text{d}_1,\text{d}_0} = (\lambda_D d_1 - \gamma_D(d_0+1)) - (\lambda_D d_1 - \gamma_D d_0) = -\gamma_D$. In Table 4.4 this is equivalent to moving between columns (i.e. d_0 to $d_0 + 1$ at a particular value of d_1).

If it is assumed that $\lambda_D = \gamma_D = \theta$, then θ is the change in $\text{ACE}_{\text{d}_1,\text{d}_0}$ as the difference in dose under the two counterfactual situations increases by one unit. Assuming constant effects of doses on potential outcomes ($\lambda = \gamma$) is reasonable under the assumption that potential dose is exchangeable between levels of random treatment allocation (assumption C5a).

Table 4.4: Grid of $\text{ACE}_{\text{d}_1,\text{d}_0}$ at levels of dose when offered control (d_0) or active intervention (d_1).

		$\mathbf{d_0}$					
		0	**1**	**2**	**3**	...	g
$\mathbf{d_1}$	**0**	0	–	–	–		–
	1	λ	0	–	–		–
	2	2λ	$2\lambda-\gamma$	0	–		–
	3	3λ	$3\lambda-\gamma$	$3\lambda-2\gamma$	0		–
	⋮						
	g	$g\lambda$	$g\lambda-\gamma$	$g\lambda-2\gamma$	$g\lambda-3\gamma$		0

New estimator for $\text{ACE}_{\text{d}_1,\text{d}_0}$ and correspondence with linear model

Observed outcome is expressed as whichever potential outcome is seen in the observable world:

$$\begin{aligned}
Y_i =&(1-R_i)Y_i(R=0)+R_iY_i(R=1)\\
=&(1-R_i)\big[Y_b+\gamma_D D_i(R=0)+\tau_i\big]\\
&+R_i\big[Y_b+\lambda_D(D_i(R=1))+\epsilon_i\big]\\
=&Y_b+\gamma_D D_i(R=0)+\tau_i-R_iY_b-R_i\gamma_D D_i(R=0)-R_i\tau_i\\
&+R_iY_b+R_i\lambda_D(D_i(R=1))+R_i\epsilon_i\\
=&Y_b+\gamma_D(1-R_i)D_i(R=0)+R_i\lambda_D(D_i(R=1))+(1-R_i)\tau_i+R_i\epsilon_i\\
=&Y_b+\gamma_D D_0+\lambda_D D_1+(1-R_i)\tau_i+R_i\epsilon_i \qquad (4.12)
\end{aligned}$$

The linear dose-response model is substituted for potential outcome between steps one and two using assumption C2. In the final step, D_0 is dose if R=0 and zero otherwise, and D_1 is dose if R=1 and zero otherwise. Assuming that $\gamma_D=\lambda_D=\theta_D$ under C5a, θ_D represents the effect of dose on outcome among compliers. Further defining $\eta_i=((1-R_i)\tau_i+R_i\epsilon_i)$, this means Equation 4.12 can be expressed simply as:

$$Y_i=Y_b+\theta_D D+\eta_i \qquad (4.13)$$

where $D=D_0+D_1$ is the observed dose under either condition. It is notable that this

equation includes no parameters which describe the effect of treatment offer (randomisation) on outcome. This is the impact of the assumption that the effect of treatment offer on outcome is conducted entirely by dose (i.e. the exclusion restriction).

An unbiased estimate of the parameter θ_D cannot be obtained using an ordinary least squares regression of Y on D. This is because the error term η is a function of D, i.e. D is correlated with η and therefore endogenous. Despite this, the effect of D on Y can be estimated because R is an instrument for D (assumptions A-I1 and A-I2). Therefore Equation 4.13 can be expressed in the following expectations:

$$\begin{aligned}\mathrm{E}[Y_i|R_i] =& Y_b + \theta_D \mathrm{E}[D_i|R_i] + \mathrm{E}[\eta_i|R_i] \\ =& Y_b + \theta_D \mathrm{E}[D_i|R_i]\end{aligned}$$

Including a vector of baseline covariates, this would be expressed as:

$$\begin{aligned}\mathrm{E}[Y_i|R_i, \mathbf{X_i}] =& Y_b + \theta_D \mathrm{E}[D_i|R_i, \mathbf{X_i}] + \mathrm{E}[\eta_i|R_i, \mathbf{X_i}] \\ =& Y_b + \theta_D \mathrm{E}[D_i|R_i, \mathbf{X_i}]\end{aligned}$$

The causal parameter of the effect of dose on treatment effect can be estimated using either the IV(2SLS) or IV(ATR) methods described earlier. The parameter estimated is the change in treatment effect associated with a one-unit increase in dose. The dose-response estimator for this causal parameter in the context of continuous dose of treatment in both the control and active intervention arms is referred to here as **E-IV8**.

Missing outcome values

This section has described four estimators: **E-IV5** (modified Bloom/ratio estimator with proportions of observed compliers, non-compliers, contaminators, and non-contaminators predicting outcome), **E-IV6** (Bloom/ratio estimator), **E-IV7** (2SLS estimator with binary treatment receipt in both arms) and **E-IV8** (2SLS with continuous measure of treatment receipt in both trial arms). In a trial in which there are missing outcome data, estimators **E-IV6**, **E-IV7**, and **E-IV8** make the strongest assumptions regarding the missing data generating mechanism. They assume that the probability of missingness is unrelated to the values of missing data (i.e. $\Pr(N|Y_o, Y_m) = \Pr(N|Y_o)$); this is assume that missing data

are MAR. Estimator **E-IV5** makes a less restrictive missingness assumption as it allows observed compliance and contamination to predict it ($\Pr(N|Y_o, Y_m, C, K) = \Pr(N|Y_o, C, K)$). Multiple imputation, as described in Section 4.4.2, can be employed to allow predictors of missingness (including observed compliance and contamination), thereby relaxing the assumptions of the estimators described in this section.

Summary

The instrumental variable methods that I have derived in this section provide a flexible approach for estimating the causal effect of treatment receipt using a semi-parametric model. They are capable of accounting for treatment non-compliance and contamination and allow manifest baseline variables to predict treatment receipt, outcome, or missingness of outcome. The causal effect of treatment receipt and its precision can be estimated using the Bloom or 2SLS estimators. In this section I have extended existing methodology for estimating the change in causal treatment effect for a one-unit increase in the difference of potential dose, where there are continuous measures of treatment receipt in both intervention and control trial arms.

4.6 Discussion

This chapter began with a statement about the challenges of conducting RCTs in mental health. Trials in this area of medicine often test complex interventions where there is scope for appreciable non-adherence with treatment regimen. It is not a strong statement to assume that the phenomenon of non-adherence is likely to be linked to drop-out from data collection and therefore estimators that address non-adherence must be capable of handling a variety of missingness assumptions.

Two types of efficacy estimand were defined. The first, ATE, is the effect of observable treatment receipt on outcome. This could be estimated using the as-treated estimator. It was shown that this estimator is unbiased only under the assumption that potential outcome is exchangeable between levels of observed treatment receipt. This is also true of the per protocol estimator, which estimates ATT. Exchangeability is a strong assumption because of the likelihood of confounding, given that these estimators throw away the protection of randomisation. The second type of efficacy estimand involves a conceptual leap into the world of potential outcomes. For binary treatment receipt, CACE is the effect of treatment amongst a subpopulation who would receive treatment

when offered it and would not when offered the comparator treatment. ACE_{d_1,d_0} was defined as the causal effect of some difference in potential dose on outcome amongst that subpopulation where potential dose is greater under offer of intervention than control.

CACE can be estimated using a stratification estimator (under the framework of principal stratification) or an IV estimator. These estimators make very similar sets of assumptions regarding the data. In mental health trials, where many treatments cannot be kept blind, the most problematic of these is the exclusion restriction. This may be particularly vulnerable in trials of interventions where it is possible that simply the offer of treatment may lead to participants altering their behaviour. If this were to happen it would open a pathway from random treatment allocation to outcome that does not go through treatment receipt (and would lead to bias).

I have described the development of a novel consistent estimator for ACE_{d_1,d_0}. This extension of the method of Maracy and Dunn (2011) allows the estimation of efficacy where there is both non-compliance and contamination. This estimator relies on similar assumptions to those when treatment receipt was a binary measure together with an assumption regarding the shape of the relationship between dose and effect. The parameter is interpreted as the causal effect of treatment associated with a one-unit increase in the difference in dose between the counterfactual worlds. It may improve interpretability if this parameter were converted into the ATE at a particular level of this difference in potential dose. For example, it could be converted into the causal effect at the maximum difference. This might be of interest where treatment receipt is a measure of therapy session attendance. This would inform patients and clinicians what they might expect to be the maximal effect of treatment.

Comparing the utility of the stratification and IV estimator types, only the former class is able to make the assumption of LI (the weakest missingness assumption). A comparison between estimators making various LI and MAR assumptions and a discussion of their relative merits is particularly important in the efficacy analysis of a trial. The other major point of note when comparing the estimator types is that currently only the IV estimators have been developed to handle continuous measures of treatment receipt. There is potentially some opportunity for likewise development of the stratification estimators.

As mentioned in Chapter 1, the randomisation-based efficacy estimators in effect swap the bias of the ITT approach (if we define this as a biased estimator of efficacy) for

variance inflation. This implies a limitation in the model for the CACE estimator: it is not identifying who the latent compliers are, it is simply using an estimate of their relative frequency. The stratification and IV estimators can be made considerably more powerful by the inclusion of predictors of treatment receipt. This has trial design implications regarding the selection of variables at baseline. For instance, it is worth considering whether to ask participants at this point how enthusiastic they are to receive treatment (such a measure would likely be predictive of latent class membership).

Chapter 5

Monte Carlo simulation study comparing two trial design options for addressing contamination

5.1 Background and aims

This thesis aims to evaluate methods for estimating efficacy in trials of complex interventions with contamination. Specifically, it aims to inform trialists and funders about the best trial design to facilitate efficacy assessment in the presence of contamination.

I have described two prominent design methods for addressing the problem of contamination in trials that target treatment efficacy. The first, which was shown to be common in the scoping review in Chapter 2, is to use cluster randomisation together with an estimator of ATE that accounts for clustering of outcome data. Provided that clusters are defined at the level at which contamination is thought to occur, this option prevents contamination by design and estimates efficacy. I refer to this as design option A throughout this chapter. The second design option is to allocate treatment randomly to individuals, accept that contamination will take place, measure treatment receipt for all participants, and then use a randomisation-based estimator of efficacy as captured by a local estimand such as CACE. I refer to this as design option B. This chapter aims to compare the statistical performance of these two approaches and therefore addresses the primary research objective of this project.

Existing theory suggests that estimator **E-ITT** under design option A is consistent for

ATE provided that selection, attrition and assessment biases can be avoided in the CRCT. Estimators of efficacy under design option B were summarised and developed in Chapter 4. In particular, estimators **E-IV7** and **E-IV8** provide consistent estimation of CACE (binary treatment receipt) and $ACE_{d_1+1,d_0} - ACE_{d_1,d_0}$ (continuous treatment receipt), respectively. In this chapter I use simulation techniques to evaluate the relative efficiency of the two designs. This builds upon the work by Keogh-Brown et al. (2007) which compared similar designs, although they did not explicitly target treatment efficacy (the research was described in Chapter 1). Their results suggested that the relative efficiency of the design options was driven by strength of clustering and amount of contamination. The results indicated that at a low strength of clustering and large amount of contamination design option A was favoured. However, the research simulated only a very small number of trial scenarios and fixed the ICC at a single level. In addition they compared sample sizes required to achieve a certain level of power rather than assessing the performance of the estimators at a range of sample sizes.

This chapter will simulate plausible trials of complex interventions in mental health using parameter values suggested by relevant datasets (the D6, CONMAN, REFOCUS and systematic assessment of care needs trials, and the results of the scoping review). The chapter targets the contamination process that was found to be most common in Chapter 2, which was clinicians being trained in both treatments under examination and then providing the active treatment to those in the control arm. The chapter is organised in two sections, a simulation substudy where contamination is measured on a binary scale and a second substudy where contamination is measured on a continuous scale (i.e. dose of treatment). The first substudy, but not the second one, will also investigate the impact on the relative efficiency of the two design options of the presence of never takers (stratum observed as non-compliers in the treatment arm). I will use simulation to evaluate the statistical properties of relevant efficacy estimators for data generated under designs A and B.

5.2 Simulation study 1: Binary measure of treatment receipt under design option B

5.2.1 Contamination process in therapy trial

The imagined scenario is a trial evaluating a new therapy in addition to treatment as usual (TAU) therapy against TAU alone. Both therapies are delivered by a set of therapists. In this first simulation substudy receipt of therapy is binary. That is to say a patient either receives the active condition ($T = 1$) or not ($T = 0$). It is not possible for patients to receive therapies other than those tested in the trial. In other words not receiving the active condition implies receipt of the control condition and *vice versa*.

Clustering

Post-randomisation outcomes of patients treated by the same therapist (therapist clusters) may be correlated as a result of the following processes:

- Cluster-level variables affect the level of outcome: the therapists have catchment areas from which their patients would be recruited and patients from the same catchment area share the cluster environment (e.g. area deprivation). In addition, characteristics of accessible patient populations might differ between such therapist clusters (e.g. the patient population potentially being treated by a female therapist may include a larger proportion of females). Thus baseline outcomes are clustered at the level of the therapist and this clustering will persist in post-treatment outcomes to some extent.

- Cluster-level variables affect the change in outcome under the control condition: it is possible that under therapist-delivered TAU, disease progression is more similar between patients who share the same therapist than between those who do not due to shared cluster environments (e.g. facilities in clinic, interaction with the same therapist) or characteristics of patient populations.

- Cluster-level variables affect the size of the intervention response: it is possible that the treatment response, that is the difference in change over time under the control and the active treatment, is more similar between patients treated by the same therapist due to shared cluster environments (e.g. clinic where treatment

takes place, interaction with the same therapist) or characteristic of the patient population.

To reflect this two-level population structure – patients are nested in therapist catchment area clusters – I use the subscript j to label the clusters and the subscript i to label the patients within the j^{th} cluster.

Non-compliance and contamination

The process whereby non-compliance or contamination is thought to take place is therapists being trained in both therapies, and the patient either not receiving a sufficient dose when offered active therapy (non-compliance with active condition $T(R=1)=0$), or receiving a sufficient dose when offered control therapy (contamination of control condition $T(R=0)=1$). Therapists' being trained in both therapies may or may not lead to participants receiving a sufficient dose. This is to say that treatment receipt varies within therapist clusters. For simplicity it is assumed that if therapists were only trained in delivering one or the other therapy then their patients would always receive the therapy that was offered to them (i.e. no contamination or non-compliance). Thus the patient population can be partitioned into four strata according to the therapy they would receive from a therapist who is trained in both conditions as follows:

- Compliers ($S=1$): Receive therapy when active condition is offered and receive control therapy when the control condition is offered $[T(R=1)-T(R=0)=1]$,
- Never takers ($S=2$): Do not receive active therapy under either offer $[T(R=1)=T(R=0)=0]$,
- Always takers ($S=3$): Receive active therapy irrespective of offer $[T(R=1)=T(R=0)=1]$,
- Defiers: Receive control therapy when active condition is offered and receive active therapy when the control condition is offered $[T(R=1)-T(R=0)=-1]$.

It is assumed that there are no defiers in the patient population and I define the following relative sizes of the subpopulations: $p_1 = \text{Prob(Complier)}$, $p_2 = \text{Prob(Never taker)}$ and $p_3 = \text{Prob(Always taker)}$, $p_1 + p_2 + p_3 = 1$.

Simulation hypotheses

I made the following predictions regarding the statistical performance of the competing design options under this contamination process:

1. A cluster randomised trial design (design option A) with therapy allocated at the level of the therapist and the therapist trained in only one of the therapies, and analysed using estimator **E-ITT** (intention-to-treat estimator) will provide an unbiased estimate of efficacy formalised by ATE.

2. An individual randomised trial design (design option B) with the therapist delivering both therapies, treatment receipt measured and analysed using the as-treated estimator will provide a biased estimate of efficacy.

3. The magnitude of the bias of the as-treated estimator (design option B) will be driven by parameters determining the strength of hidden confounding.

4. An individual randomised trial design (design option B) with the therapist delivering both therapies, treatment receipt measured and analysed using estimator **E-IV7** (two stage least squares estimator) will provide an asymptotically unbiased estimate of efficacy formalised by CACE.

5. The relative efficiency of the two competing approaches will be driven by parameters describing the population cluster structure and those determining the strength of respective instrumental variables.

5.2.2 Data generating models

The simulation study mimicked the clustered structure of the target population before any trial took place, and simulated observed post-treatment outcomes under two trial design options:

A. Cluster randomisation at the level of the therapist with the therapist only trained in one of the competing therapies.

B. Individual level randomisation with therapists being trained and delivering both competing therapies; measurement of binary therapy receipt for each participant.

Data resulting from either trial design were then analysed using the respective estimation approach.

I proceeded in the following steps:

1. Simulation of the distribution of outcome in the target population before any intervention took place (baseline outcome);

2. Simulation of four potential post-intervention outcomes in the target population:

   - Outcome that would be observed if patients were offered the control condition (i.e. $R = 0$) and were treated by a therapist who is only trained in that therapy (i.e. $Q = 0$), that is I simulated $Y(R = 0, Q = 0)$,
   - Outcome that would be observed if patients were offered the control condition and were treated by a therapist who is trained in both therapies (i.e. $Q = 1$), that is I simulated $Y(R = 0, Q = 1)$,
   - Outcome that would be observed if patients were offered the active condition and were treated by a therapist who is only trained in that therapy, that is I simulated $Y(R = 1, Q = 0)$,
   - Outcome that would be observed if patients were offered the active condition and were treated by a therapist who is trained in both therapies, that is I simulated $Y(R = 1, Q = 1)$;

   This allowed me to define individual therapy offer effects (IREs) under both treatment delivery options as the contrasts:

   - If the treatment was delivered as planned under trial design A: $\text{IRE}_{\text{A}} := Y_{ij}(R = 1, Q = 0) - Y_{ij}(R = 0, Q = 0)$,
   - If the treatment was delivered as planned under trial design B: $\text{IRE}_{\text{B}} := Y_{ij}(R = 1, Q = 1) - Y_{ij}(R = 0, Q = 1)$;

3. Simulation of the patient sampling and treatment allocation implied by trial design options A and B, and mapping of potential outcomes onto observed Y under the design options;

4. For each data set generated under design options A or B respectively I calculated the proposed estimator and its SE. I repeated the process to determine the sampling distributions of the estimators.

Step 1) – Baseline outcome

Baseline outcome $Y_{0,ij}$ had a hierarchical structure as it was made up of a random variable that varied at the level of the individual, ε_{ij}, and another random variable that varied at the level of the therapist cluster, v_j. For convenience I scaled $\text{var}(Y_{0,ij}) = 1$.

I generated $Y_{0,ij} := e_{ij} + v_j \quad$ with $e_{ij} \sim N(0, 1-\rho)$ and independently $v_j \sim N(0, \rho)$.

Here parameter $\rho \in [0,1]$ denotes the intraclass correlation coefficient (ICC1) measuring the proportion of variance in Y_0 that was due to factors that varied at the therapist catchment area level.

Step 2) – Potential post-intervention outcomes

To start with I generated the three strata of patients according to the therapies they would receive from a therapist who is trained in both conditions ($Q = 1$):

$S_{ij} \sim \text{Mu}[(p_1, p_2, 1-p_1-p_2), 1]$

It then followed that under design option B,

$$T_{ij}(R=0) = \begin{cases} 0 \text{ if } S_{ij} = 1 \text{ or } S_{ij} = 2 \\ 1 \text{ if } S_{ij} = 3 \end{cases} \quad \text{and } T_{ij}(R=1) = \begin{cases} 0 \text{ if } S_{ij} = 2 \\ 1 \text{ if } S_{ij} = 1 \text{ or } S_{ij} = 3 \end{cases}$$

I then generated the four potential post-treatment outcomes.

Potential outcome $Y_{ij}(R=0, Q=0)$:
This is the outcome that would be observed if patients were offered the control condition and were treated by a therapist who is only trained in that therapy and is given by:

$Y_{ij}(R=0, Q=0) := \alpha Y_{0,ij} + \varepsilon_{ij} + \omega_j \quad$ with $\varepsilon_{ij}|S = s \sim N[\mu_s, 1-\alpha^2-\xi-g]$ and independently $\omega_j \sim N(0, \xi)$.

For simplicity it was assumed that the mean did not change over time under the control condition, i.e. $\text{E}[Y_{ij}(R=0, Q=0)] = 0$. Parameter $\alpha \in [0,1]$ was the effect of baseline on post-intervention outcome.

Parameter $\xi \in [0,1]$ denotes the intraclass correlation coefficient (ICC2) measuring the proportion of variance in baseline-adjusted $Y_{ij}(R=0, Q=0) - \alpha Y_{0,ij}$ that is due to factors that varied at the therapist level.

Random variable ε_{ij} represents variables that explain variability in change over the treatment period. Such variables can include post-randomisation variables and can thus be affected by (pre-randomisation) stratum membership. Thus I allowed their mean to depend on stratum membership, i.e. $\mathrm{E}(\varepsilon_{ij}|S=s)=\mu_s$ with $\mathrm{E}(\varepsilon_{ij})=p_1\mu_1+p_2\mu_2+p_3\mu_3=0$. Noting that $\mathrm{var}(\varepsilon_{ij})=\mathrm{E}[\mathrm{var}(\varepsilon_{ij}|S)]+\mathrm{var}[\mathrm{E}(\varepsilon_{ij}|S)]=\mathrm{var}(\varepsilon_{ij}|S)+\mathrm{var}[\mu_S]=\mathrm{var}(\varepsilon_{ij}|S)+p_1\mu_1^2+p_2\mu_2^2+p_3\mu_3^2$, it is stated $g:=p_1\mu_1^2+p_2\mu_2^2+p_3\mu_3^2$. It was assumed that the outcome variance was not increased under the control condition, i.e. $\mathrm{var}[Y_{ij}(R=0,Q=0)]=\mathrm{var}(Y_{0,ij})=1$. And to ensure that this held I set $\mathrm{var}(\varepsilon_{ij})=\mathrm{var}(\varepsilon_{ij}|S)+g=1-\alpha^2-\xi$.

Potential outcome $Y_{ij}(R=0,Q=1)$:
This is the outcome that would be observed if patients were offered the control condition and were treated by a therapist who is trained in both therapies. It can be affected by treatment that is actually received in this situation, $T(R=0)$.

I generated $Y_{ij}(R=0,Q=1):=\alpha Y_{0,ij}+\beta T_{ij}(R=0)+\varepsilon_{ij}+\omega_j$.

Here parameter β represents the effect of receiving the therapy. Such contamination can increase the variance in the presence of always takers since $\mathrm{var}[Y_{ij}(R=0,Q=1)]=\mathrm{var}[Y_{ij}(R=0,Q=0)]+\mathrm{var}[\beta T_{ij}(R=0)]+2\beta\mathrm{cov}[\varepsilon_{ij},T_{ij}(R=0)]=1+\beta^2p_3(1-p_3)+2\beta p_3(1-p_3)\big[\mu_3-\big(\frac{p_1\mu_1+p_2\mu_2}{p_1+p_2}\big)\big]=1+[\beta^2+2\beta\{\mu_3-\big(\frac{p_1\mu_1+p_2\mu_2}{p_1+p_2}\big)\}]p_3(1-p_3)$.

The expression for the covariance between the error term and treatment receipt under offer of control was found as follows:

$$\begin{aligned}\mathrm{cov}[\varepsilon_{ij},T_{ij}(R=0)]&=\mathrm{E}[\varepsilon_{ij}T_{ij}(R=0)]-\mathrm{E}[\varepsilon_{ij}]\mathrm{E}[T_{ij}(R=0)]\\&=p_3\mathrm{E}[\varepsilon_{ij}|T_{ij}(R=0)=1]-\big[p_3\mathrm{E}[\varepsilon_{ij}|T_{ij}(R=0)=1]+\\&\qquad(1-p_3)\mathrm{E}[\varepsilon_{ij}|T_{ij}(R=0)=0]\big]p_3\\&=p_3(1-p_3)\big[\mathrm{E}[\varepsilon_{ij}|T_{ij}(R=0)=1]-\mathrm{E}[\varepsilon_{ij}|T_{ij}(R=0)=0]\big]\\&=p_3(1-p_3)\left(\mu_3-\left(\frac{p_1\mu_1+p_2\mu_2}{p_1+p_2}\right)\right)\end{aligned}$$

Potential outcome $Y_{ij}(R=1,Q=0)$:
Next is the outcome that would be observed if patients were offered the active condition and were treated by a therapist who is only trained in that therapy. This is given by

$$Y_{ij}(R=1,Q=0):=\alpha Y_{0,ij}+\gamma+\varepsilon_{ij}+\omega_j+\tau_{ij}(Q=0)\quad\text{with}\quad\tau_{ij}(Q=0)\sim N(0,\vartheta)$$

Here parameter $\gamma = \mathrm{E}\{\mathrm{IRE_A}\}$, i.e. this parameter was the causal effect of treatment receipt, ATE. The added error term $\tau_{ij}(Q=0)$ introduced treatment effect heterogeneity in that $\mathrm{var}\{\mathrm{IRE_A}\} = \mathrm{var}\{\tau_{ij}(Q=0)\} = \vartheta$. Such treatment effect heterogeneity is allowed to increase the variance of the post-treatment outcome under active condition compared to that under the control condition; specifically $\mathrm{var}[Y_{ij}(R=1,Q=0)] = 1+\vartheta$. (For simplicity I assumed that this latest error term did not include therapist effects.)

Potential outcome $Y_{ij}(R=1,Q=1)$**:**
Finally, the outcome that would be observed if patients were offered the active condition and were treated by a therapist who is trained in both therapies. This can also be affected by the treatment that was actually received in this situation, $T(R=1)$. The potential outcome was modelled,

$$Y_{ij}(R=1,Q=1) := \alpha Y_{0,ij} + \beta T_{ij}(R=1) + \varepsilon_{ij} + \omega_j + \tau_{ij}(Q=1)$$

Here parameter β again represents the effect of receiving the active condition (it models non-compliance). The error term $\tau_{ij}(Q=1)$ also represents treatment effect heterogeneity, but this time under treatment delivery by therapists who are trained in both therapies ($Q=1$). Under such treatment delivery the population can contain never takers ($S=2$) and always takers ($S=3$) as well as compliers ($S=1$). It was assumed that the expected error term within a stratum was zero; i.e. $\mathrm{E}\big[\tau_{ij}(Q=1)|S=s\big]=0$, that is random treatment effect variability within a stratum. This implies two exclusion restriction assumptions as follows for never takers, $\mathrm{E}\big\{Y_{ij}(R=1,Q=1) - Y_{ij}(R=0,Q=1)|S=2\big\} = \mathrm{E}\big\{\tau_{ij}(Q=1)|S=2\big\} = 0$, and for always takers, $\mathrm{E}\big\{Y_{ij}(R=1,Q=1) - Y_{ij}(R=0,Q=1)|S=3\big\} = \mathrm{E}\big\{\tau_{ij}(Q=1)|S=3\big\} = 0$. In words, the offer of treatment is assumed not to have a (mean) effect on outcome for never takers or always takers. From this also follows that $\mathrm{CACE} = \mathrm{E}\big\{Y_{ij}(R=1,Q=1) - Y_{ij}(R=0,Q=1)|S=1\big\} = \beta + \mathrm{E}\big\{\tau_{ij}(Q=1)|S=1\big\} = \beta$.

Variability in mean treatment effects across strata contributes to treatment effect heterogeneity. Specifically $\mathrm{var}\big[\mathrm{E}\{\mathrm{IRE_B}|S\}\big] = p_1(\beta - p_1\beta)^2 + p_2(0-p_1\beta)^2 + p_3(0-p_1\beta)^2 = (1-p_1)p_1\beta^2$ (proof below). It was assumed that an individual's treatment effect relative to the stratum mean does not depend on the therapist's training and this ensured that the total treatment effect heterogeneity was the same for both delivery options, that is $\mathrm{var}[\mathrm{IRE_B}] = \mathrm{var}[\mathrm{IRE_A}] = \vartheta$. To this end I define $\tau_{ij}(Q=1) := w\tau_{ij}(Q=0)$ with $w := \big\{1-(1-p_1)p_1\beta^2/\vartheta\big\}^{0.5}$.

The variance of the expectation of IRE_B conditioning on S was found as follows:

$$\begin{aligned}
\text{var}\big[\text{E}\{\text{IRE}_\text{B}|S\}\big] &= p_1(\beta - p_1\beta)^2 + p_2(0 - p_1\beta)^2 + p_3(0 - p_1\beta)^2 \\
&= p_1(\beta^2 - 2p_1\beta^2 + p_1^2\beta^2) + p_2(p_1^2\beta^2) + p_3(p_1^2\beta^2) \\
&= p_1\beta^2 - 2p_1^2\beta^2 + p_1^3\beta^2 + p_2p_1^2\beta^2 + p_3p_1^2\beta^2 \\
&= \beta^2(p_1 - 2p_1^2 + p_1^3 + p_2p_1^2 + p_3p_1^2) \\
&= p_1\beta^2(1 - 2p_1 + p_1^2 + p_2p_1 + p_3p_1) \\
&= p_1\beta^2(1 - 2p_1 + p_1(p_1 + p_2 + p_3)) \\
&= p_1\beta^2(1 - 2p_1 + p_1) = p_1\beta^2(1 - p_1)
\end{aligned}$$

Step 3) – Observable trial data

Next, the generation of trial data was mimicked under either trial design. Let n denote the trial sample size and k the number of patients recruited from each therapist catchment area. A trial sample was generated of n patients and m therapists ($k = n/m$ patients per therapist catchment area) by first randomly sampling m units at the therapist level and then randomly sampling k patients for each of the selected catchment areas. Under either design baseline outcome $Y_{0,ij}$ was measured for $i = 1, ..., k; j = 1, ..., m$.

Under trial design option A treatment offer was randomly allocated at the level of the therapist with 50:50 allocation ratio captured by binary variable $R_{(A),j}$ (with levels 1="active", 0="control"). The allocation ratio was fixed at exactly 50:50 (or as close as possible to this if there was an odd number of clusters). This led to therapists being nested within trial arms. Observed post-randomisation outcomes could then be generated by mapping values onto respective potential outcomes under design A:

$$Y_{(A),ij} := R_{(A),j}Y_{ij}(R = 1, Q = 0) + [1 - R_{(A),j}]Y_{ij}(R = 0, Q = 0)$$

Note that post-randomisation outcome $Y_{(A),ij}$ has a hierarchical structure due to incorporating two clustered variables: (i) the baseline levels of the outcome and (ii) change in the outcome under the control condition (natural progression).

The set of variables available for analysis under trial design A was then given by: $Y_{0,ij}$, $R_{(A),j}$ and $Y_{(A),ij}$.

In contrast, under trial design option B treatment offer was randomly generated at the patient level. Stratified randomisation (here the stratum is the therapist) was used to ensure that half the patients of a therapist received either treatment offer (in 50:50 allocation ratio). This allocation was captured by variable $R_{(B),ij}$ (with levels 1="active", 0="control"). This type of randomisation led to therapists being crossed with trial arms (and consequently having to be trained in both therapies). Again, clustered observed post-randomisation outcomes were generated by mapping values onto respective potential outcomes under design B:

$$Y_{(B),ij} := R_{(B),ij}Y_{ij}(R=1,Q=1) + [1-R_{(B),ij}]Y_{ij}(R=0,Q=1)$$

Design option (B) further stipulated that binary treatment receipt T (with levels 1="received" and 0="not received") was measured. Thus to produce this observable variable, potential treatment receipt was mapped as such:

$$T_{ij} = R_{(B),ij}T_{ij}(R=1) + [1-R_{(B),ij}]T_{ij}(R=0)$$

The set of variables available for analysis under trial design (B) was then given by: $Y_{0,ij}$, $R_{(B),ij}$, T_{ij} and $Y_{(B),ij}$.

Step 4) – Efficacy estimators

Design A: ITT maximum likelihood estimator with random intercept for cluster (therapist). This estimator, without the random effect, was referred to in the estimation methods chapter as estimator **E-ITT** (Section 4.3.1). The estimator included baseline outcome as a covariate in order to reflect the fact that this was strongly related to outcome at follow-up. A random intercept was needed in order to allow for correlations within therapist clusters (this approach was described in Section 4.4.2).

Design B1: As-treated maximum likelihood estimator of outcome on treatment receipt with random effect for cluster (therapist). This estimator, without the random effect, was described in Section 4.3.2. The estimator included baseline outcome as a covariate in order to reflect the fact that this was strongly related to outcome at follow-up. A random intercept was needed in order to allow for correlations within therapist clusters (this approach was described in Section 4.4.2).

Design B2: Generalised 2SLS estimator with random effect for cluster (therapist). This

model, without the random effect, was referred to previously as estimator **E-IV7** (Section 4.5.2). The estimator included baseline outcome as a covariate in both stages in order to reflect the fact that this was strongly related to outcome at follow-up. A random intercept was needed in order to allow for correlations within therapist clusters (this approach was described in Section 4.4.2).

5.2.3 Simulation study design

Seven simulation parameters were varied and three were held constant. I investigated all combinations (4,752 trial scenarios) of these parameters and used 1000 iterations per scenario. A list of input parameters and their levels is given in Table 5.1.

Table 5.1: Summary of input levels of simulation parameters when simulating a binary measure of treatment receipt.

Parameter	Description	Levels of parameter
Sample size (n)		100; 200; 500; 1000
Treatment effect: β	CACE (design option B)	0.2; 0.5; 0.8
Treatment effect: γ	ATE (design option A)	0.2; 0.5; 0.8
ICC2 (ξ)	Level 2 clustering; due to variance at level of therapist	0.01; 0.02; 0.05; 0.10
Size of clusters (k)		5; 10; 20
$\mu_2 - \mu_2$ and $\mu_3 - \mu_1$	Confounding due to non-adherence	0.2; 0.5; 0.8
Strength of instrument (p_1)	This is also the proportion of latent compliers	0.4; 0.5; 0.6; 0.7; 0.8; 0.9
Proportion of never takers (p_2)	This is also the proportion of non-compliers within the active intervention arm	0; 0.2
α	Effect of outcome at baseline on outcome at follow-up	0.7
ICC1 (ρ)	Level 2 clustering; due to variance at level of catchment area	0.1
Treatment heterogeneity (ϑ)	This is the variance of IRE_A	0.2

Number of trial scenarios: 4 sample sizes, 3 treatment effect sizes, 4 intraclass correlation coefficients (ICC2), 3 cluster sizes, 3 magnitudes of confounding, 6 strengths of instrument with no treatment non-compliance, 5 strengths of instrument with a typical amount of treatment non-compliance, i.e. 4,752 trial scenarios.

Those parameters that were varied were sample size (to study estimators' asymptotic behaviour), treatment effect size, ICC2, cluster size, confounding bias due to non-adherence

(design option B), strength of IV (design option B), and proportion of never takers (design option B). Confounding bias was varied in order to assess its predicted impact on the as-treated estimator (design option B1; this was simulation hypothesis 3). ICC2, cluster size and strength of IV were varied in order to assess their anticipated impact on the relative efficiency of the design options A and B2 (simulation hypothesis 5). I selected two levels of the proportion of never takers in order to investigate whether there was any impact of the presence of non-compliance on the relative efficiency of these design options. The levels that I chose for these parameters were as follows.

- Sample size was varied (n=100, 200, 500, 1000). These levels reflected the range of sample sizes that were observed in the systematic review (median 186, IQR 100-372). This choice also reflected the range of sample sizes in the four mental health trial datasets that partly motivated this research (Marshall et al., 2004; Weaver et al., 2014; Slade et al., 2015; Ismail et al., 2018). The range of sample sizes in these trials was 210 to 403.

- Two target parameters: β (CACE) and γ (ATE). For simplicity ATE was assumed to be the same as CACE, irrespective of whether the therapist was trained in one or both treatments. Three levels of standardised effect sizes (β and γ) were selected: small (0.2), medium (0.5) and large (0.8).

- Strength of clustering in outcome: Clustering was driven by ICC1 (see below) and ICC2 for the data generating models. A realistic range of levels for ICC2 was chosen. This was based on the results of the systematic review and four mental health trial datasets research (Marshall et al., 2004; Weaver et al., 2014; Slade et al., 2015; Ismail et al., 2018). The systematic review found a median ICC of 0.05 and interquartile range of 0.03-0.09. The ICCs for primary outcomes in the motivating datasets ranged from <0.001 to 0.1. Based on this and the systematic review, four levels of ICC were selected: 0.01, 0.02, 0.05, 0.1.

- Cluster size: the levels of cluster size were chosen to represent realistic trials of complex interventions within the context of mental health. The systematic review suggested a median of 10 and interquartile range of 6-27. Mean cluster sizes in the four motivating mental health trial datasets ranged from 4 to 18 (Marshall et al., 2004; Weaver et al., 2014; Slade et al., 2015; Ismail et al., 2018). These numbers were rounded to some extent in order for all clusters to be complete at

all the levels of sample size listed earlier. Therefore three levels were chosen: 5, 10, 20.

- Hidden confounding of the treatment effect under design option B was driven by μ's (stratum effect on treatment-free outcome). I considered $\mu_2 - \mu_1$ "strength of confounding due to non-compliance with active treatment offer" and $\mu_3 - \mu_1$ "strength of confounding due to contamination of the control treatment offer". Since they represented standardised effects of S, I chose small ($d = 0.2$), moderate ($d = 0.5$) and large ($d = 0.8$) differences.

- Strength of IV R under design option B: the IV assumptions for R were implemented by all of the effect of R on Y being through T, by R being unrelated to unobserved confounders, and by assuming there were no defiers when constructing $T(R = 1)$ and $T(R = 0)$. The strength of R as an IV was measured by p_1 (proportion of latent compliers). Parameter p_1 was varied from 0.4 to 0.9, in steps of 0.1. This was consistent with the results of the systematic review which found a median proportion of the control group receiving intervention (contamination) of 0.13 (IQR 0.05-0.33), with a few outlying values around 0.6. When there was some non-compliance (see below), the maximum strength of IV was adjusted so that it, added to the proportion of non-compliance, did not rise above one.

- Proportion of never takers under design option B: the proportion of never takers was also the proportion of those within the active intervention arm who did not receive intervention (i.e. non-compliers). I chose two levels of p_2: 0, 0.2. I.e. no non-compliance (which ensured that design options A and B were comparable) and a typical level of it (see DiMatteo, 2004; Nose et al., 2003).

Those parameters that were held constant were the effect of outcome at baseline on outcome at follow-up, ICC1 and the treatment heterogeneity. These were set as follows.

- Effect of outcome at baseline on outcome at follow-up: α. This parameter was fixed at a level that implied that outcome measured at baseline and follow-up were strongly related. This parameter was set to 0.7.

- Clustering due to ICC1 could be removed by the estimators, which covaried on baseline outcome. ICC1 was set to 0.1.

- Treatment heterogeneity: Note that in the plan described earlier, the variance of the IRE_B and IRE_A were constrained to be the same (i.e. ϑ). This was achieved by $w := \left\{1-(1-p_1)p_1\beta^2/\vartheta\right\}^{0.5}$, where w was a factor affecting the variance of $\tau_{ij}(Q=0)$. This implied a restriction that $(1-p_1)p_1\beta^2 < \vartheta$. Theta was set to 0.2 as this was the lowest (round) level that could be sustained given the levels of p_1 and β.

Sampling distribution of estimates:

For each trial scenario, output was saved from the 1000 simulated datasets. The following output was saved for each trial scenario in order to assess the sampling distribution of the estimates:

- Mean of the estimates,
- Median,
- SD of the mean (i.e. the true SE for the sampling disribution; also known as the empirical SE (EmpSE)),
- Minimum,
- Lower 2.5 percentile,
- 25th percentile,
- 75th percentile,
- Upper 2.5 percentile,
- Maximum,
- Proportion of simulations where associated 95% CI includes true efficacy estimand ("nominal confidence level").

Properties of estimator:

The following output was saved in order to explore properties of the estimators:

- Bias $= \text{E}[\hat{\gamma}]-\gamma$ for option A; $\text{E}[\hat{\beta}]-\beta$ for options B1 and B2,
- Mean SE of estimator (i.e. the estimated SE),
- Mean t-statistic,

- Monte Carlo standard error (MCSE) $= \frac{\text{EmpSE}}{\sqrt{2(n_{\text{sim}}-1)}}$, where n_{sim} is the number of simulation iterations.

Comparison of estimators:

In order to calculate the relative efficiency of the estimators under design options A and B2, I generated the ratio of their SEs:

- Ratio of SEs = SE from option A / SE from option B2.

5.2.4 Simulation findings

I have split the results from the simulations into three parts: an investigation of the bias of the estimators under designs options A, B1 and B2, an assessment of the model-based SE compared to true SE (for the unbiased estimators), and an evaluation of the relative efficiency of the unbiased estimators using the true SEs. The assessment of bias provides results for simulation hypotheses 1-4, and the relative efficiency of the unbiased estimators gives results for hypothesis 5.

Bias

The results of the Monte Carlo simulations showed that the ITT estimator (named earlier **E-ITT**), which was the estimator of efficacy under design option A, was an unbiased estimator of ATE (parameter γ in these data simulations). Mean absolute bias of this estimator in the scenario in which there was no non-compliance in the intervention arm is shown in Table 5.2. The table summarises bias to three decimal places by levels of sample size and strength of confounding. It shows that there was negligible bias at all levels of these variables.

Table 5.2: Absolute bias of ITT estimator (option A) rounded to three decimal places, with binary treatment receipt and no non-compliance in the intervention arm. Results were averaged over cluster sizes and magnitudes of treatment effect sizes.

Sample size	**Strength of confounding**		
	0.2	0.5	0.8
100	0.000	−0.001	0.000
200	0.000	0.000	0.000
500	0.000	0.000	0.000
1000	0.000	0.000	0.000

Mean absolute bias of the as-treated approach as an estimator of efficacy, which was part of design option B1, is displayed in Table 5.3. The results are summarised by sample size, strength of confounding, and proportion of compliers. There was considerable bias at every level of these variables. The amount of bias appeared to be related to strength of confounding and proportion of compliers, with the most when there was large amounts of confounding and when the proportion of compliers was low.

Table 5.3: Absolute bias of as-treated estimator (option B1) rounded to three decimal places, with binary treatment receipt and no non-compliance in the intervention arm. Results were averaged over ICCs, cluster sizes and magnitudes of treatment effect sizes.

Sample size;	**Proportion of compliers**					
strength of confounding	0.4	0.5	0.6	0.7	0.8	0.9
100						
0.2	0.150	0.134	0.115	0.093	0.068	0.038
0.5	0.377	0.337	0.289	0.232	0.171	0.093
0.8	0.606	0.538	0.462	0.374	0.273	0.149
200						
0.2	0.151	0.134	0.117	0.094	0.068	0.037
0.5	0.378	0.336	0.289	0.234	0.170	0.094
0.8	0.604	0.538	0.463	0.376	0.272	0.149
500						
0.2	0.151	0.134	0.116	0.094	0.069	0.037
0.5	0.378	0.336	0.289	0.235	0.170	0.093
0.8	0.604	0.539	0.463	0.376	0.272	0.150
1000						
0.2	0.152	0.135	0.116	0.094	0.068	0.037
0.5	0.378	0.337	0.290	0.235	0.171	0.094
0.8	0.605	0.539	0.463	0.376	0.273	0.150

Mean absolute bias of the IV estimator (named earlier **E-IV7**) as an estimator of efficacy (CACE, parameter β in this chapter), which was part of design option B2, is shown in Table 5.4. Results are summarised by sample size, strength of confounding, and proportion of compliers. Bias was very small at all levels of these variables. This demonstrated that it was consistent for all levels of confounding.

The as-treated estimator was dropped because of the clear evidence of bias. The asymptotically unbiased estimators (**E-ITT** and **E-IV7**) were investigated further. The next section will compare the model-based SE with the true SE for these two estimators.

Table 5.4: Absolute bias of IV estimator (option B2) rounded to three decimal places, with binary treatment receipt and no non-compliance in the intervention arm. Results were averaged over ICCs, cluster sizes and magnitudes of treatment effect sizes.

Sample size; strength of confounding	**Proportion of compliers**					
	0.4	0.5	0.6	0.7	0.8	0.9
100						
0.2	−0.004	−0.003	−0.002	−0.002	−0.001	0.000
0.5	−0.009	−0.007	−0.001	−0.004	−0.001	0.000
0.8	−0.012	−0.009	−0.007	−0.004	−0.001	−0.001
200						
0.2	−0.002	−0.002	0.001	−0.001	0.000	−0.001
0.5	−0.004	−0.004	−0.002	−0.002	0.000	0.001
0.8	−0.007	−0.007	−0.002	−0.001	−0.002	−0.001
500						
0.2	0.000	0.000	0.000	0.000	0.001	0.000
0.5	−0.002	−0.002	−0.001	−0.001	−0.001	0.000
0.8	−0.005	−0.002	−0.002	−0.001	−0.001	0.000
1000						
0.2	0.001	0.000	−0.001	0.000	0.000	0.000
0.5	−0.002	0.000	0.000	−0.001	0.000	0.000
0.8	−0.002	−0.001	−0.001	−0.001	−0.001	0.000

SE estimation

Summaries of the ratio between mean model-based SE and the true SE (the standard deviation of the point estimates) showed that the two were very similar for estimator **E-ITT**. At the smallest sample size of 100 participants, the model-based SE was roughly 10% smaller than the true SE. As sample size increased this ratio approached one. In other words the model-based estimate of the SE was correct for estimator **E-ITT** for large sample sizes. The ratio was unrelated to strength of confounding. Summaries of the ratio, which are displayed by sample size and strength of confounding, are given in Table 5.5.

Summaries of the ratio between mean model-based SE and true SE showed that estimator **E-IV7** performed well. At the smallest sample size the model-based SE was a slight underestimate of the true SE, but not to the same extent as for estimator **E-ITT**. As sample size increased, so did the ratio, suggesting that the SE was consistent. However, at the largest sample size the model-based SE was overestimated by roughly 3%. There was little evidence of a relationship between this ratio and strength of confounding or proportion of compliers. Summaries of this ratio, displayed by sample size, strength of

Table 5.5: Model SE / true SE of ITT estimator (option A) rounded to three decimal places, with binary treatment receipt and no non-compliance in the intervention arm. Results were averaged over cluster sizes and magnitudes of treatment effect sizes.

Sample size	Strength of confounding		
	0.2	0.5	0.8
100	0.899	0.901	0.897
200	0.946	0.947	0.944
500	0.978	0.974	0.975
1000	0.987	0.988	0.989

confounding and proportion of compliers, are shown in Table 5.6.

Table 5.6: Model SE / true SE of IV estimator (option B2) rounded to three decimal places, with binary treatment receipt and no non-compliance in the intervention arm. Results were averaged over ICCs, cluster sizes and magnitudes of treatment effect sizes.

Sample size; strength of confounding	Proportion of compliers					
	0.4	0.5	0.6	0.7	0.8	0.9
100						
0.2	0.984	0.989	0.986	0.989	0.984	0.976
0.5	0.978	0.999	0.985	0.987	0.988	0.980
0.8	0.984	0.994	0.993	0.990	0.990	0.978
200						
0.2	1.005	1.004	1.005	1.003	1.004	1.001
0.5	1.001	1.005	1.000	0.999	1.005	0.999
0.8	1.005	1.005	0.999	0.998	1.000	1.005
500						
0.2	1.014	1.011	1.014	1.021	1.017	1.016
0.5	1.024	1.014	1.019	1.015	1.017	1.023
0.8	1.015	1.016	1.026	1.016	1.022	1.016
1000						
0.2	1.034	1.025	1.023	1.024	1.029	1.020
0.5	1.026	1.034	1.027	1.025	1.025	1.031
0.8	1.026	1.028	1.027	1.022	1.017	1.016

Monte Carlo standard error

Summaries of the MCSE under design option A (estimator **E-ITT**) were always less than 0.01. I found that MCSE decreased with increasing sample size. Summaries of the MCSE, which are displayed by sample size, are given in Table 5.7. The MCSE was small in comparison to the estimator's standard error.

Table 5.7: Monte Carlo standard error for ITT estimator (option A) rounded to three decimal places, with binary treatment receipt and no non-compliance in the intervention arm.

Sample size	
100	0.004
200	0.003
500	0.002
1000	0.001

Summaries of the MCSE under design option B2 (estimator **E-IV7**) were always less than 0.01. I found that MCSE decreased with increasing sample size and with increasing proportion of compliers. Summaries of the MCSE, which are displayed by sample size and proportion of compliers, are given in Table 5.8. The MCSE was small in comparison to the estimator's standard error.

Table 5.8: Monte Carlo standard error for IV estimator (option B2) rounded to three decimal places, with binary treatment receipt and no non-compliance in the intervention arm.

Sample size	**Proportion of compliers**					
	0.4	0.5	0.6	0.7	0.8	0.9
100	0.009	0.007	0.006	0.005	0.004	0.004
200	0.006	0.005	0.004	0.003	0.003	0.003
500	0.004	0.003	0.003	0.002	0.002	0.002
1000	0.003	0.002	0.002	0.002	0.001	0.001

Relative efficiency

The efficiency of estimator **E-ITT** (design option A) relative to **E-IV7** (design option B2) is displayed in Table 5.9 for the scenario in which there was no treatment non-compliance and in Table 5.10 for the scenario in which there was contamination and non-compliance (20% of intervention arm did not receive treatment). Mean relative efficiency is summarised by sample size, proportion of compliers, ICC, and cluster size. Cells where the ratio was greater than one are shaded in grey. These cells indicate the instances in which the SE of **E-ITT** is greater than that of **E-IV7**, i.e. the estimation of efficacy under design option B2 is more precise.

When there was no non-compliance, in general the results showed that the estimation of efficacy under design option B2 was more precise than that under design option A when the proportion of compliers was large, cluster size was high, and ICC was large. The simulated trial scenarios provided a large range of efficiency ratios, from 0.397 to 2.062. When the amount of contamination was very large (60%, i.e. proportion of compliers of 0.4) design option A was favoured at every simulated level of cluster size and ICC. At 50% and 40% contamination, design option A was almost always favoured, with the exception of the largest levels of cluster size and ICC. When the amount of contamination was 30%, design option A was favoured at most levels of cluster size and ICC, except at the highest level of ICC. The picture was more mixed at 20% contamination. When the amount of contamination was 10%, design option B2 was favoured at most levels of cluster size and ICC. The exceptions to this were when ICC was very low (0.01) and when both ICC was moderately low (0.02) and cluster size was also low.

When strength of clustering was low (small cluster size, or ICCs of 0.01 or 0.02), design option A was often favoured, apart from in some instances where the amount of contamination was small. At moderate levels of either cluster size (10) or ICC (0.02 or 0.05) and when contamination was 30% or under, design option B2 was favoured in roughly half the scenarios. For combinations of moderate to large cluster sizes (10 or 20) and ICCs (0.05 or 0.1), design option B was usually favoured when contamination was 30% or under. Within levels of proportion of compliers, an increase in cluster size was associated with a shift in the efficiency ratio towards option B (i.e. greater ratio). The pattern of the increasing ratio became increasingly pronounced as ICC increased.

When there was some treatment non-compliance, the pattern was similar when comparing the relative efficiencies of the two design options. The difference was that the proportion of compliers was capped at 0.8, which was now indicative of no treatment contamination. The meant that all the descriptions above were now true at 20 percentage points of contamination lower than previously described. For example, it was stated previously that at 30% contamination, design option A was favoured at most levels of cluster size and ICC, except at the highest level of ICC. With 20% non-compliance, this statement would be true at 10% contamination.

Table 5.9: Relative efficiency of design options A and B2, with no non-compliance in the intervention arm. Cells represent ratios of mean estimated standard errors of design option A divided by design option B2. The results are summarised by sample sizes and proportions of compliers (rows), and intraclass correlation coefficients and cluster sizes (k) (columns). Results were averaged over magnitudes of effect size and strength of confounding.

Sample size; proportion of compliers	**ICC=0.01**			**ICC=0.02**			**ICC=0.05**			**ICC=0.1**		
	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$
100												
0.4	0.397	0.421	0.481	0.419	0.444	0.536	0.450	0.539	0.688	0.544	0.676	0.881
0.5	0.508	0.538	0.614	0.531	0.571	0.680	0.579	0.684	0.882	0.682	0.876	1.148
0.6	0.610	0.661	0.720	0.645	0.717	0.828	0.709	0.831	1.053	0.844	1.047	1.381
0.7	0.720	0.776	0.882	0.743	0.825	0.935	0.823	0.981	1.223	0.977	1.224	1.614
0.8	0.817	0.871	0.962	0.846	0.961	1.100	0.943	1.133	1.376	1.102	1.419	1.857
0.9	0.917	0.980	1.086	0.935	1.041	1.197	1.042	1.227	1.551	1.232	1.571	2.062
200												
0.4	0.407	0.438	0.475	0.420	0.476	0.531	0.466	0.558	0.671	0.565	0.681	0.893
0.5	0.518	0.543	0.601	0.541	0.594	0.663	0.597	0.699	0.840	0.725	0.885	1.124
0.6	0.635	0.669	0.712	0.654	0.709	0.798	0.712	0.825	1.023	0.822	1.033	1.338
0.7	0.733	0.760	0.847	0.744	0.839	0.928	0.837	0.971	1.188	0.978	1.218	1.574
0.8	0.809	0.897	0.939	0.849	0.962	1.051	0.967	1.129	1.354	1.129	1.411	1.824
0.9	0.911	0.979	1.060	0.941	1.058	1.180	1.066	1.242	1.495	1.242	1.577	2.006
500												
0.4	0.408	0.441	0.479	0.421	0.480	0.531	0.478	0.566	0.682	0.553	0.709	0.889
0.5	0.512	0.557	0.597	0.531	0.598	0.662	0.593	0.694	0.869	0.706	0.894	1.138
0.6	0.631	0.669	0.727	0.649	0.720	0.799	0.718	0.842	1.014	0.851	1.051	1.352
0.7	0.715	0.774	0.843	0.754	0.833	0.930	0.836	1.009	1.184	0.984	1.258	1.606
0.8	0.811	0.888	0.951	0.857	0.939	1.054	0.961	1.123	1.376	1.101	1.414	1.806
0.9	0.920	0.992	1.063	0.948	1.042	1.174	1.061	1.262	1.522	1.248	1.573	2.041
1000												
0.4	0.415	0.443	0.480	0.431	0.466	0.541	0.479	0.576	0.680	0.568	0.715	0.927
0.5	0.523	0.561	0.597	0.542	0.598	0.679	0.599	0.706	0.873	0.706	0.894	1.150
0.6	0.617	0.669	0.714	0.653	0.709	0.801	0.725	0.856	1.039	0.859	1.074	1.359
0.7	0.720	0.782	0.825	0.755	0.848	0.926	0.833	0.977	1.214	0.972	1.274	1.565
0.8	0.835	0.878	0.947	0.849	0.936	1.053	0.949	1.144	1.355	1.122	1.404	1.796
0.9	0.919	0.975	1.085	0.960	1.060	1.173	1.065	1.226	1.519	1.210	1.548	2.000

Table 5.10: Relative efficiency of design options A and B2, with some non-compliance in the intervention arm. Cells represent ratios of mean estimated standard errors of design option A divided by design option B2. The results are summarised by sample sizes and proportions of compliers (rows), and intraclass correlation coefficients and cluster sizes (k) (columns). Results were averaged over magnitudes of effect size and strength of confounding.

Sample size; proportion of compliers	**ICC=0.01**			**ICC=0.02**			**ICC=0.05**			**ICC=0.1**		
	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$
100												
0.4	0.372	0.397	0.470	0.385	0.447	0.530	0.437	0.499	0.663	0.511	0.655	0.879
0.5	0.505	0.538	0.602	0.516	0.565	0.691	0.575	0.684	0.844	0.672	0.842	1.135
0.6	0.607	0.645	0.737	0.630	0.694	0.825	0.703	0.840	1.042	0.818	1.044	1.357
0.7	0.728	0.770	0.859	0.738	0.831	0.961	0.832	0.981	1.224	0.964	1.223	1.618
0.8	0.806	0.872	0.972	0.845	0.922	1.080	0.948	1.108	1.383	1.097	1.387	1.853
200												
0.4	0.392	0.434	0.464	0.414	0.464	0.519	0.459	0.548	0.666	0.554	0.682	0.867
0.5	0.520	0.550	0.596	0.518	0.586	0.669	0.601	0.707	0.844	0.701	0.876	1.115
0.6	0.628	0.666	0.714	0.652	0.697	0.789	0.705	0.838	1.034	0.838	1.078	1.348
0.7	0.718	0.773	0.847	0.753	0.833	0.932	0.848	0.991	1.206	0.989	1.221	1.548
0.8	0.817	0.869	0.940	0.845	0.935	1.067	0.942	1.122	1.359	1.121	1.399	1.770
500												
0.4	0.406	0.431	0.476	0.421	0.467	0.538	0.480	0.562	0.669	0.564	0.699	0.905
0.5	0.523	0.554	0.589	0.538	0.598	0.658	0.600	0.704	0.846	0.694	0.884	1.119
0.6	0.620	0.655	0.727	0.645	0.727	0.808	0.734	0.845	1.037	0.850	1.072	1.357
0.7	0.713	0.778	0.846	0.746	0.830	0.934	0.835	0.986	1.219	0.989	1.254	1.612
0.8	0.815	0.884	0.965	0.849	0.946	1.067	0.937	1.107	1.362	1.105	1.404	1.817
1000												
0.4	0.411	0.447	0.478	0.432	0.484	0.525	0.478	0.568	0.687	0.563	0.707	0.922
0.5	0.523	0.552	0.610	0.536	0.584	0.657	0.602	0.713	0.856	0.707	0.882	1.134
0.6	0.629	0.668	0.728	0.643	0.711	0.807	0.733	0.865	1.035	0.859	1.064	1.377
0.7	0.743	0.785	0.831	0.745	0.822	0.940	0.851	0.991	1.194	0.987	1.233	1.592
0.8	0.837	0.874	0.937	0.842	0.955	1.049	0.953	1.136	1.361	1.128	1.396	1.832

5.3 Simulation study 2: Continuous measure of treatment receipt under design option B

5.3.1 Contamination process in therapy trial

The scenario is similar to that described earlier in that I am imagining a trial that investigates the effect of therapy in addition to TAU therapy compared to TAU alone; therapies are delivered by a therapist. The difference is that treatment receipt is now a continuous measure. Therefore a patient receives some dose of active therapy ($D > 0$) or not ($D = 0$), or they receive a fixed dose of TAU (control) therapy. It is assumed that this therapy has a full dose (which I call m). As before, it is not possible for patients to receive therapies other than those tested in the trial.

Clustering:

As before, post-randomisation outcomes of patients treated by the same therapist (therapist clusters) may be correlated as a result of the following processes:

- Cluster-level variables that affect the level of outcome,
- Cluster-level variables that affect the change in outcome under the control condition,
- Cluster-level variables that affect the size of the intervention response.

Non-compliance and contamination

The process by which non-compliance and contamination are thought to take place is the same as in simulation substudy 1: therapists being trained in both therapies leading to non-receipt of active therapy amongst those offered it and receipt of active therapy amongst those offered the TAU therapy. The difference is that receipt of active therapy is now defined as a continuous measure (dose). More specifically, it is on a ratio scale because zero indicates an entire lack of treatment receipt. Non-compliance with active therapy takes place when $D(R = 1) < m$ and contamination of the control condition occurs when $D(R = 0) > 0$. Once again it is assumed that if therapists were trained in only one therapy then their patients would always receive the therapy that was offered to them (i.e. $D(R = 1) = m$ and $D(R = 0) = 0$). Generalising definitions from binary to continuous treatment receipt, the patient population can be divided into strata according

to the therapy they would receive from a therapist who is trained in both conditions as follows:

- Dose responders: receive a greater dose of active therapy when active condition is offered than they do when control condition is offered $[D(R=1)-D(R=0)>0]$,
- Dose never takers: receive no dose of active therapy when either active or control conditions are offered $[D(R=1)=D(R=0)=0]$,
- Dose always takers: receive the same, non-zero dose of active therapy when either active or control conditions are offered $[D(R=1)=D(R=0)>0]$
- Dose defiers: receive a smaller dose of active therapy when active condition is offered than they do when control condition is offered $[D(R=1)-D(R=0)<0]$.

It is assumed that there are no dose defiers in the patient population and define $q_1 =$ Prob(Dose complier), $q_2 =$ Prob(Dose never taker), $q_3 =$ Prob(Dose always taker); $q_1 + q_2 + q_3 = 1$.

Simulation hypotheses:

1. A cluster randomised trial design with therapy allocated at the level of the therapist and the therapist trained in only one of the therapies, and analysed using estimator **E-ITT** (intention-to-treat estimator) will provide an unbiased estimate of efficacy formalised by ATE.
2. An individual randomised trial design with the therapist delivering both therapies, treatment receipt measured, and analysed using the as-treated estimator will provide a biased estimate of efficacy as formalised by ATE.
3. The magnitude of absolute bias of the as-treated estimator will be driven by parameters determining the strength of hidden confounding.
4. An individual randomised trial design with the therapist delivering both therapies, treatment receipt measured, and analysed using estimator **E-IV8** (two stage least squares dose-response estimator) will provide an asymptotically unbiased estimate of efficacy formalised by $\text{ACE}_{\text{m},0}$.
5. The relative efficiency of the two competing unbiased approaches will be driven by parameters describing the population cluster structure and those determining the strength of respective instrumental variables.

5.3.2 Data generating models

The simulation study mimicked a clustered structure in the target population (before trial took place) and simulated post-treatment outcomes under the two design options in a manner similar to before. The difference was that in design option B there was a continuous measure of treatment receipt for each participant.

I proceeded following the same four steps.

Step 1) – Baseline outcome

This was the same as in simulation substudy 1. Specifically, I generated outcome at baseline with level-one and level-two random effects: $Y_{0,ij} := e_{ij} + v_j$ with $e_{ij} \sim N(0, 1-\rho)$ and independently $v_j \sim N(0, \rho)$.

Step 2) – Potential post-intervention outcomes

Dose of active therapy under offer of control therapy, $D_{ij}(R=0)$:
To start with I generated potential dose under the offer of control according to the therapy patients would receive from a therapist who is trained in both conditions ($Q=1$). All values in this distribution must be non-negative, with a possible peak at zero representing those who received no dose of active intervention under the offer of control (if dose never takers were present in the population).

In order to generate potential dose under offer of control I created two separate random variables. These variables allowed the zero-inflation caused by the possible presence of dose never takers and the distribution of dose on a continuous scale for dose compliers and dose always takers. The product of the two variables represents potential dose of active therapy under offer of control.

First, I sampled from a binomial distribution, where zero values represent dose compliers or dose never takers and ones indicate dose compliers or dose always takers:

$A_{ij}(R=0)|q_3 > 0 \sim \text{bin}(1, 1-q_4)$, with $q_4 := \frac{q_2}{(q_2+q_3)}$. The fraction $\frac{q_2}{(q_2+q_3)}$ represents the proportion who receive a potential treatment dose of zero of those who do not adhere to treatment offer. This was done in order to preserve the ratio of dose never takers to dose always takers as determined by $q_2 : q_3$. When there is no treatment receipt under offer of control (i.e. $q_3 = 0$), each value of treatment receipt is zero and therefore:

$$\left[A_{ij}(R=0)|q_3=0\right]=0$$

For full dose $m_D > 0$, where m_D represents the full dose of active therapy a participant could receive when offered control and where $m_D \leq m$, I sampled from a uniform distribution in order to obtain dose under offer of control (for dose compliers and dose always takers).

$$B_{ij}(R=0)|m_D>0 \sim \text{unif}(0, m_D)$$

No treatment receipt under offer of control could arise when maximum dose under offer of control is set to zero:

$$\left[B_{ij}(R=0)|m_D=0\right]=0$$

Potential dose amongst those who are offered control is then:

$$D_{ij}(R=0) := A_{ij}(R=0)B_{ij}(R=0)$$

Difference in dose of active therapy between trial arms, F_{ij}:
I then generated a binary random variable (S^*) that represented latent compliance (where 1=dose compliers and 0=dose never takers and dose always takers). The requirement for this variable was to preserve the proportions of dose compliers amongst those who received no dose under offer of control and amongst those who received some dose under offer of control. For simplicity I assumed that the ratio of dose compliers between these two groups was equal to the ratio of dose never takers to dose always takers (i.e. q_4). When dose of active therapy under offer of control was zero, I generated the distribution of S^* as follows,

$$S^*|D_{ij}(R=0)=0 \sim \text{bin}\left(1, \frac{q_1 q_4}{(q_1 q_4+q_2)}\right)$$

Note that $\frac{q_1 q_4}{q_1 q_4+q_2} = \frac{\frac{q_1 q_2}{(q_2+q_3)}}{\frac{q_1 q_2}{(q_2+q_3)}+q_2} = \frac{q_1 q_2}{q_1 q_2+(q_2+q_3)q_2} = \frac{q_1 q_2}{q_1 q_2+(1-q_1)q_2} = \frac{q_1 q_2}{q_2} = q_1.$

When dose of active therapy under offer of control was positive, I generated the distribution of S^* as follows,

$$S^*|D_{ij}(R=0)>0 \sim \text{bin}\left(1, \frac{q_1(1-q_4)}{q_1(1-q_4)+q_3}\right)$$

Note that $\frac{q_1(1-q_4)}{q_1(1-q_4)+q_3} = \frac{q_1-\frac{q_1 q_2}{(q_2+q_3)}}{q_1-\frac{q_1 q_2}{(q_2+q_3)}+q_3} = \frac{q_1(q_2+q_3)-q_1 q_2}{q_1(q_2+q_3)-q_1 q_2+q_3(q_2+q_3)} = \frac{q_1 q_2+q_1 q_3-q_1 q_2}{q_1 q_2+q_1 q_3-q_1 q_2+q_3(1-q_1)}$
$= \frac{q_1 q_3}{q_1 q_3+q_3-q_1 q_3} = \frac{q_1 q_3}{q_3} = q_1.$

The difference in potential dose under the offers of control and treatment is denoted by random variable F where $F_{ij} := D_{ij}(R=1) - D_{ij}(R=0)$. I generated F_{ij} rather than $D_{ij}(R=1)$ because of the requirement that there are no dose defiers in the population (i.e. $D_{ij}(R=1) \geq D_{ij}(R=0)$). I allowed for this requirement by generating F_{ij} so that all values were non-negative. I wanted the flexibility to vary the minimum level of F_{ij} and I refer to this minimum as n_F, where $0 \leq n_F \leq m$. In order to prevent dose of active therapy under offer of this treatment being greater than full dose m, the uniform distribution that F_{ij} was sampled from had an upper limit that was bounded by the dose of active therapy under offer of control.

F_{ij} is conditional on the level of S^*. The distribution of F_{ij} for dose compliers when the maximum difference in potential dose under the offers of control and treatment was greater than the minimum difference was,

$$F_{ij}|S^* = 1, n_F < m \sim \text{unif}\big(n_F, m - D(R=0)\big)$$

For the extreme case where all dose compliers receive the maximum difference in dose of active therapy between the trial arms, this implies that $n_F = m$. When these parameters were set in such a way, F_{ij} was fixed at the level of the maximum difference.

$$\big[F_{ij}|S^* = 1, n_F = m\big] = m$$

Dose never takers and dose always takers receive the same dose of active therapy under offers of control and treatment,

$$\big[F_{ij}|S^* = 0\big] = 0$$

A separate way of viewing the generation of F_{ij} is to proceed in a similar manner to the generation of $D_{ij}(R=0)$, as described earlier in this section. Say that F_{ij} is the product of two random variables, called J_{ij} and H_{ij}, where J_{ij} enables zero inflation (for possible presence of dose never takers and dose always takers) and H_{ij} represents the measure of treatment dose (for the difference in potential dose amongst the dose compliers). These variables are distributed in the following manner:

$J_{ij} \sim \text{bin}(1, q_1)$
$H_{ij} \sim \text{unif}(n_F, m - D_{ij}(R=0)$

I have described the generation of F_{ij} in this alternate way because this greatly simplifies matters later on when finding an expression for the variance of F_{ij}.

Dose of active therapy under offer of active therapy, $D_{ij}(R=1)$:
Having generated both dose of active therapy under offer of control therapy and the difference in potential dose under the offers of control and active therapies, I then generated potential dose under offer of treatment, $D_{ij}(R=1) := D_{ij}(R=0) + F_{ij}$.

Potential outcome $Y_{ij}(R=0,Q=0)$:
Next, I generated the four potential post-treatment outcomes. The outcome that would be observed if patients were offered the control condition and were treated by a therapist who is only trained in that therapy is given by

$Y_{ij}(R=0,Q=0) := \alpha Y_{0,ij} + \varepsilon_{ij} + \omega_j$ with $\varepsilon_{ij}|D(R=0)=d_0, F=f \sim N[\mu_{d_0 f}, 1-\alpha^2-\xi-g]$ and independently $\omega_j \sim N(0,\xi)$.

As a reminder, ε_{ij} is a level-one random effect and represents patient-level heterogeneity during follow-up. ω_j is a level-two random effect and represents therapist-level heterogeneity during follow-up, and ξ is the variance component (it is also ICC2).

Random variable ε_{ij} represents variables that can explain variability in change over the treatment period. Its mean, conditional on the levels of treatment receipt under the offer of control and the difference in potential dose, is defined as follows. $\mathrm{E}[\varepsilon_{ij}|D(R=0)=d_0, F=f] = \mu_{d_0 f}$, with $\mathrm{E}[\varepsilon_{ij}] = -\Delta(d_0 - \mathrm{E}[D_{ij}(R=0)]) + \Delta(f - \mathrm{E}[F]) = 0$. Δ is the confounding effect of either receiving some dose of treatment when offered control or of non-compliance with treatment when offered active intervention. It is the parameter that represents the change in the (standardised) level-one error term for Y associated with a one-unit increase in either potential dose under offer of control or the difference in potential dose between the offers of treatment and control. The different signs before Δ, depending on whether the explanatory variable is $D(R=0)$ or F, reflect the fact that the confounding effects of greater contamination (greater d_0) and non-compliance (smaller f) are in the same direction.

It was assumed that the variance of outcome under control was not increased by the presence of these population strata, i.e. $\mathrm{var}[Y_{ij}(R=0,Q=0)] = \mathrm{var}(Y_{0,ij}) = 1$. And to ensure that this held I set $\mathrm{var}(\varepsilon_{ij}) = \mathrm{var}(\varepsilon_{ij}|D(R=0)=d_0, F=f) + g = 1-\alpha^2-\xi$. Note that, using the Law of Total Variance, $\mathrm{var}(\varepsilon_{ij}) = \mathrm{E}_{(D(0),F)}\big[\mathrm{var}_\varepsilon(\varepsilon|D(0),F)\big] + \mathrm{E}_{D(0)}\Big\{\mathrm{var}_F\big[\mathrm{E}_\epsilon(\varepsilon|D(0),F)|D(0)\big]\Big\} + \mathrm{var}_{D(0)}\big[\mathrm{E}_\varepsilon(\varepsilon|D(0))\big] = \mathrm{var}(\varepsilon_{ij}|D(0)=d_0, F=f) + g$, i.e. $g := \mathrm{E}_{D(0)}\Big\{\mathrm{var}_F\big[\mathrm{E}_\epsilon(\varepsilon|D(0),F)|D(0)\big]\Big\} + \mathrm{var}_{D(0)}\big[\mathrm{E}_\varepsilon(\varepsilon|D(0))\big]$. These terms for g

could be expressed in terms of knowable quantities. Starting by expressing the first term for g in such terms and letting $D_{ij}(R=0) = D(0)$ and $F_{ij} = F$,

$$
\begin{aligned}
&\mathrm{E}_{D(0)}\Big\{\mathrm{var}_F\big[\mathrm{E}_\epsilon(\varepsilon|D(0),F)|D(0)\big]\Big\} \\
&\qquad = \mathrm{E}_{D(0)}\Big\{\mathrm{var}_F\big[-\Delta(d_0 - \mathrm{E}[D(0)]) + \Delta(F - \mathrm{E}(F)|D(0) = d_0\big]\Big\} \\
&\qquad = \mathrm{E}_{D(0)}\Big\{\mathrm{var}_F\big[\Delta(F - \mathrm{E}(F)|D(0) = d_0\big]\Big\} \\
&\qquad = \mathrm{E}_{D(0)}\Big\{\mathrm{var}_F\big[\Delta F|D(0) = d_0\big]\Big\} \\
&\qquad = \Delta^2\mathrm{E}_{D(0)}\Big\{\mathrm{var}_F\big[F|D(0) = d_0\big]\Big\} \\
&\qquad = \Delta^2\mathrm{E}_{D(0)}\Big\{\mathrm{E}_J\Big[\mathrm{var}_F\big[F|D(0) = d_0, J = j\big]\Big]\Big\} \\
&\qquad = \Delta^2\mathrm{E}_{D(0)}\Big\{\mathrm{Prob}(J=0)\ \mathrm{var}_F\big[F|D(0) = d_0, J = 0\big] \\
&\qquad\qquad + \mathrm{Prob}(J=1)\ \mathrm{var}_F\big[F|D(0) = d_0, J = 1\big]\Big\} \\
&\qquad = \Delta^2\mathrm{E}_{D(0)}\Big\{q_1\mathrm{var}_F\big[F|D(0) = d_0, J = 1\big]\Big\} \\
&\qquad = \Delta^2\mathrm{E}_{D(0)}\Big\{q_1\frac{1}{12}(m - D(0) - n_F)^2\Big\} \\
&\qquad = \frac{\Delta^2 q_1}{12}\ \mathrm{E}_{D(0)}\big\{(m - D(0) - n_F)^2\big\} \\
&\qquad = \frac{\Delta^2 q_1}{12}\ \mathrm{E}_{D(0)}\big\{m^2 - 2D(0)m + (D(0))^2 - 2mn_F + 2D(0)n_F + n_F^2\big\} \\
&\qquad = \frac{\Delta^2 q_1}{12}\ \big\{m^2 - 2\ \mathrm{E}[D(0)]m + \mathrm{E}[(D(0))^2] - 2mn_F + 2\ \mathrm{E}[D(0)]n_F + n_F^2\big\} \\
&\qquad = \frac{\Delta^2 q_1}{12}\ \big\{m^2 - 2\ \mathrm{E}[D(0)]m + \mathrm{var}[D(0)] + \mathrm{E}[D(0)]^2 - 2mn_F \\
&\qquad\qquad\qquad + 2\ \mathrm{E}[D(0)]n_F + n_F^2\big\}
\end{aligned}
$$

Then, the second term for g:

$$
\begin{aligned}
&\mathrm{var}_{D(0)}\Big[\mathrm{E}_{\varepsilon}(\varepsilon|D(0))\Big] \\
&\quad= \mathrm{var}_{D(0)}\Big\{\mathrm{E}_{F}\Big[\mathrm{E}_{\varepsilon}(\varepsilon|D(0),F)\Big]\Big\} \\
&\quad= \mathrm{var}_{D(0)}\Big\{\mathrm{E}_{F}\Big[-\Delta\big(d_0-\mathrm{E}[D(0)]\big)+\Delta\big(F-\mathrm{E}(F)\big)|D(0)=d_0\Big]\Big\} \\
&\quad= \mathrm{var}_{D(0)}\Big\{\mathrm{E}_{F}\Big[-\Delta\big(d_0-\mathrm{E}[D(0)]\big)|D(0)=d_0\Big]+\mathrm{E}_{F}\Big[\Delta\big(F-\mathrm{E}(F)\big)|D(0)=d_0\Big]\Big\} \\
&\quad= \mathrm{var}_{D(0)}\Big\{\Big[-\Delta D(0)+\Delta\mathrm{E}[D(0)]\Big]+\mathrm{E}_{F}\Big[\Delta F|D(0)=d_0\Big]-\Delta\mathrm{E}(F)\Big\} \\
&\quad= \mathrm{var}_{D(0)}\Big\{\Big[-\Delta D(0)+\Delta\mathrm{E}[D(0)]\Big]+\Delta\mathrm{E}_{J}\Big[\mathrm{E}_{F}[F|D(0)=d_0,J=j]\Big]-\Delta\mathrm{E}(F)\Big\} \\
&\quad= \mathrm{var}_{D(0)}\Big\{\Big[-\Delta D(0)+\Delta\mathrm{E}[D(0)]\Big]+\Delta\Big[(1-q_1)\mathrm{E}_{F}[F|D(0)=d_0,J=0] \\
&\qquad\qquad +q_1\mathrm{E}_{F}[F|D(0)=d_0,J=1]\Big]-\Delta\mathrm{E}(F)\Big\} \\
&\quad= \mathrm{var}_{D(0)}\Big\{\Big[-\Delta D(0)+\Delta\mathrm{E}[D(0)]\Big]+\Delta\Big[q_1\mathrm{E}_{F}[F|D(0)=d_0,J=1]\Big]-\Delta\mathrm{E}(F)\Big\} \\
&\quad= \mathrm{var}_{D(0)}\Big\{\Big[-\Delta D(0)+\Delta\mathrm{E}[D(0)]\Big]+\Delta\Big[q_1\frac{(m-d_0+n_F)}{2}\Big]-\Delta\mathrm{E}(F)\Big\} \\
&\quad= \mathrm{var}_{D(0)}\Big\{-\Delta\Big[D(0)-\mathrm{E}[D(0)]-\frac{q_1(m-d_0+n_F)}{2}+\mathrm{E}(F)\Big]\Big\} \\
&\quad= \mathrm{var}_{D(0)}\left\{-\Delta\Big[\mathrm{E}(F)-\mathrm{E}[D(0)]-\frac{q_1(m+n_F)}{2}+\Big(1+\frac{q_1}{2}\Big)D(0)\Big]\right\} \\
&\quad= \mathrm{var}_{D(0)}\left\{-\Delta\Big[\Big(1+\frac{q_1}{2}\Big)D(0)\Big]\right\} \\
&\quad= \Delta^2\Big(1+\frac{q_1}{2}\Big)^2\mathrm{var}[D(0)]
\end{aligned}
$$

In order to express these terms for g, I need expressions for the expectation of $D(0)$ and the variance of $D(0)$. They can be found as follows.

Expectation of $D_{ij}(0)$:

I now demonstrate how I found an expression for the expectation of $D_{ij}(R=0)$. As a reminder and letting $A_{ij}(R=0)=A$ and $B_{ij}(R=0)=B$, $D(0)$ is the product of A (random variable for the balance between those patients who receive some dose of active therapy under offer of control and those who receive zero dose) and B (random variable for dose of active therapy), where $A\sim\mathrm{bin}\big(1,\frac{q_3}{(q_2+q_3)}\big)$ and $B\sim\mathrm{unif}(0,m_D)$.

$$
\mathrm{E}[D(0)]=\mathrm{E}(AB)=\mathrm{E}(A)\mathrm{E}(B)=\frac{q_3}{(q_2+q_3)}\cdot\frac{m_D}{2}=\frac{m_Dq_3}{2(q_2+q_3)}
$$

Variance of $D_{ij}(0)$:

The expression for the variance of $D(0)$ can be found as follows.

$$\begin{aligned}\text{var}[D(0)] &= \text{var}(A)\text{var}(B) + \text{var}(A)\big(\text{E}(B)\big)^2 + \text{var}(B)\big(\text{E}(A)\big)^2 \\ &= \left(\frac{q_3}{(q_2+q_3)}\right)\left(1-\frac{q_3}{(q_2+q_3)}\right)\left(\frac{m_D^2}{12}\right) + \left(\frac{q_3}{(q_2+q_3)}\right)\left(1-\frac{q_3}{(q_2+q_3)}\right)\left(\frac{m_D}{2}\right)^2 \\ &\qquad + \left(\frac{m_D^2}{12}\right)\left(\frac{q_3}{(q_2+q_3)}\right) \\ &= \frac{m_D^2(4q_2q_3+q_3^2)}{12(q_2+q_3)^2}\end{aligned}$$

Expression for g:

Putting these terms together in an expression for g,

$$\begin{aligned}g &:= \frac{\Delta^2 q_1}{12}\left\{m^2 - 2\,\text{E}[D(0)]m + \text{var}[D(0)] + \text{E}[D(0)]^2 - 2mn_F + 2\,\text{E}[D(0)]n_F + n_F^2\right\} \\ &\quad + \Delta^2\Big(1+\frac{q_1}{2}\Big)^2\text{var}[D(0)] \\ &= \frac{\Delta^2 q_1}{12}\left\{m^2 - m\left(\frac{m_D q_3}{(q_2+q_3)}\right) + \left\{\frac{m_D^2(4q_2q_3+q_3^2)}{12(q_2+q_3)^2}\right\} + \left(\frac{m_D q_3}{2(q_2+q_3)}\right)^2\right. \\ &\quad \left. - 2mn_F + n_F\left(\frac{m_D q_3}{(q_2+q_3)}\right) + n_F^2\right\} + \Delta^2\Big(1+\frac{q_1}{2}\Big)^2\left\{\frac{m_D^2(4q_2q_3+q_3^2)}{12(q_2+q_3)^2}\right\}\end{aligned}$$

Potential outcome $Y_{ij}(R=0,Q=1)$:

I then generated the potential outcome that would be observed if patients were offered the control condition and were treated by a therapist who is trained in both therapies:

$$Y_{ij}(R=0,Q=1) := \alpha Y_{0,ij} + \beta D_{ij}(R=0) + \varepsilon_{ij} + \omega_j$$

Here parameter β represents the effect of receiving some dose of the active therapy on outcome. Specifically, β measures the effect of dose of active therapy (when offered control therapy). Note that β is $\text{ACE}_{\text{d}_1,\text{d}_0} - \text{ACE}_{\text{d}_1,\text{d}_0+1}$, i.e. it is the change in the causal parameter associated with a one-unit change in dose between the counterfactual worlds. Such contamination can increase the variance in the presence of dose always takers since $\text{var}[Y_{ij}(R=0,Q=1)] = \text{var}[Y_{ij}(R=0,Q=0)] + \text{var}[\beta D_{ij}(R=0)] + 2\beta\text{cov}[\varepsilon_{ij}, D_{ij}(R=0)] = 1 + \beta^2\Big(\frac{m_D^2(4q_2q_3+q_3^2)}{(12(q_2+q_3)^2}\Big) + 2\beta\Big(\text{E}[\varepsilon_{ij}D_{ij}(R=0)]\Big)$.

Potential outcome $Y_{ij}(R=1,Q=0)$:

Next, the potential outcome that would be observed if patients were offered the active condition and were treated by a therapist who is only trained in that therapy is given by,

$$Y_{ij}(R=1,Q=0) := \alpha Y_{0,ij} + \gamma + \varepsilon_{ij} + \omega_j + \tau_{ij}(Q=0) \quad \text{with} \quad \tau_{ij}(Q=0) \sim N(0,\vartheta)$$

Parameter $\gamma = \mathrm{E}[\mathrm{IRE_A}]$, i.e. this parameter is the causal effect of receiving full dose, or ATE. The added error term $\tau_{ij}(Q=0)$ introduced treatment effect heterogeneity in that $\mathrm{var}[\mathrm{IRE_A}] = \mathrm{var}[\tau_{ij}(Q=0)] = \vartheta$. Such treatment effect heterogeneity was allowed to increase the variance of the post-treatment outcome under active condition compared to that under the control condition; specifically $\mathrm{var}[Y_{ij}(R=1,Q=0)] = 1+\vartheta$. (For simplicity it was assumed that this latest error term did not include therapist effects.)

Potential outcome $Y_{ij}(R=1,Q=1)$:

Finally, I generated the potential outcome that would be observed if patients were offered the active condition and were treated by a therapist who is trained in both therapies. This potential outcome can also be affected by the dose of treatment that is actually received in this situation $D(R=1)$. The potential outcome was modelled,

$$Y_{ij}(R=1,Q=1) := \alpha Y_{0,ij} + \beta D_{ij}(R=1) + \varepsilon_{ij} + \omega_j + \tau_{ij}(Q=1)$$

The parameter β represents the effect of receiving a one-unit increase in dose of active condition on outcome. Note that this is the same parameter that was used in the construction of $Y_{ij}(R=0,Q=1)$. This is because the effect of some dose of active therapy under offer of it on outcome is equal to the effect of it under offer of control therapy. The error term $\tau_{ij}(Q=1)$ also represents treatment effect heterogeneity, but this time under treatment delivery by therapists who are trained in both therapies ($Q=1$). As previously, it was assumed that the expected error term within every strata (defined by the difference in potential dose) was zero; i.e. $\mathrm{E}\big[\tau_{ij}(Q=1)|F=f\big]=0$, that is random treatment effect variability within a stratum. This implied two exclusion restriction assumptions as follows for dose never takers, $\mathrm{E}\big[Y_{ij}(R=1,Q=1)-Y_{ij}(R=0,Q=1)|D(R=1)=D(R=0)=0\big] = \mathrm{E}\big[\tau_{ij}(Q=1)|D(R=1)=D(R=0)=0\big]=0$, and for dose always takers, $\mathrm{E}\big[Y_{ij}(R=1,Q=1)-Y_{ij}(R=0,Q=1)|D(R=1)=D(R=0)>0\big] = \mathrm{E}\big[\tau_{ij}(Q=1)|D(R=1)=D(R=0)>0\big]=0$. In words, the offer of treatment was assumed not to have a (mean) effect on outcome for dose never takers or dose always takers.

Variability in mean treatment effects within the dose compliers and across strata (dose compliers, dose always takers, and dose never takers) contributed to treatment effect

heterogeneity. Specifically, $\text{var}\Big[\text{E}\{\text{IRE}_\text{B}|F\}\Big] = \text{var}(\beta F) = \beta^2\text{var}(F)$. It was assumed that an individual's treatment effect relative to the stratum mean did not depend on the therapist's training and this ensured that the total treatment effect heterogeneity was the same for both delivery options, that is $\text{var}[\text{IRE}_\text{B}] = \text{var}[\text{IRE}_\text{A}] = \vartheta$. To this end I defined $\tau_{ij}(Q=1) := w\tau_{ij}(Q=0)$, with $w := (1-\frac{v}{\vartheta})^{0.5}$, where $v = \beta^2\text{var}(F)$. Expression v could be expressed in terms of known quantities as follows. Let $F_{ij} = F$, $J_{ij} = J$, and $H_{ij} = H$. As described earlier, random variable F is the product of J and H. J has variance $q_1(1-q_1)$. I will demonstrate how to find the expectation and variance of H, and then finally give the variance of F.

Expectation of H_{ij}:

Using the Law of Total Expectation it is possible to find an expression for the expectation of H,

$$\begin{aligned}
\text{E}[H] &= \text{E}_{D(0)}\Big[\text{E}_H(H|D(0))\Big] = \text{E}_{D(0)}\Big[\frac{1}{2}(n_F + m - D(0))\Big] \\
&= \Big[\frac{1}{2}\Big(n_F + m - \frac{m_D q_3}{2(q_2+q_3)}\Big)\Big] = \Big[\frac{n_F}{2} + \frac{m}{2} - \frac{m_D q_3}{4(q_2+q_3)}\Big] \\
&= \Big[\frac{4(n_F+m)(q_2+q_3) - 2m_D q_3}{8(q_2+q_3)}\Big] \\
&= \Big[\frac{2(n_F+m)(q_2+q_3) - m_D q_3}{4(q_2+q_3)}\Big]
\end{aligned}$$

given that $m_D > 0$, $m > n_F$, and $n_F \leq m - m_D$.

Variance of H_{ij}:

The variance of H could be found as follows, initially using the Law of Total Variance:

$$
\begin{aligned}
\mathrm{var}(H) &= \mathrm{E}[\mathrm{var}(H|D(0))] + \mathrm{var}[\mathrm{E}(H|D(0))] \\
&= \mathrm{E}\left[\frac{1}{12}(m - D(0) - n_F)^2\right] + \mathrm{var}\left[\frac{1}{2}(m - D(0) + n_F)\right] \\
&= \mathrm{E}\left[\frac{1}{12}(m^2 - 2\,\mathrm{E}[D(0)]m + \mathrm{E}[(D(0))^2] - 2mn_F + 2\,\mathrm{E}[D(0)]n_F + n_F^2)\right] \\
&\qquad + \left(\frac{1}{2}\right)^2 \mathrm{var}[D(0)] \\
&= \frac{m^2}{12} - \frac{2m\mathrm{E}[D(0)]}{12} + \frac{\mathrm{E}[(D(0))^2]}{12} - \frac{2mn_F}{12} + \frac{2n_F\mathrm{E}[D(0)]}{12} + \frac{n_F^2}{12} + \left(\frac{1}{2}\right)^2 \mathrm{var}[D(0)]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{m^2}{12} - \frac{2mm_Dq_3}{24(q_2+q_3)} + \frac{\mathrm{var[D(0)]}}{12} + \frac{\mathrm{E[D(0)]}^2}{12} - \frac{2mn_F}{12} + \frac{2n_Fm_Dq_3}{24(q_2+q_3)} + \frac{n_F^2}{12} \\
&\qquad + \frac{1}{4}\left(\frac{m_D^2(4q_2q_3+q_3^2)}{12(q_2+q_3)^2}\right) \\
&= \frac{m^2}{12} - \frac{2mm_Dq_3}{24(q_2+q_3)} + \frac{m_D^2(4q_2q_3+q_3^2)}{144(q_2+q_3)^2} + \frac{1}{12}\left(\frac{m_Dq_3}{2(q_2+q_3)}\right)^2 + \frac{2mn_F}{12} \\
&\qquad + \frac{2n_Fm_Dq_3}{24(q_2+q_3)} + \frac{n_F^2}{12} + \frac{1}{4}\left(\frac{m_D^2(4q_2q_3+q_3^2)}{12(q_2+q_3)^2}\right) \\
&= \frac{\begin{aligned}
&12m^2q_2^2 - 24mn_Fq_2^2 + 12n_F^2q_2^2 + 16m_D^2q_2q_3 - 12m_Dmq_2q_3 + 24m^2q_2q_3 \\
&\quad + 12m_Dn_Fq_2q_3 - 48mn_Fq_2q_3 + 24n_F^2q_2q_3 + 7m_D^2q_3^2 - 12m_Dmq_3^2 \\
&\quad + 12m^2q_3^2 + 12m_Dn_Fq_3^2 - 24mn_Fq_3^2 + 12n_F^2q_3^2
\end{aligned}}{144(q_2+q_3)^2}
\end{aligned}
$$

given that $m_D > 0$, $m > n_F$, and $n_F \leq m - m_D$.

Variance of F_{ij}:

This allows me to express the variance of F in terms of the expectations and variances of J and H:

$$
\begin{aligned}
\operatorname{var}(F) = \operatorname{var}(J)\operatorname{var}(H) + \operatorname{var}(J)\mathrm{E}(H)^2 + \operatorname{var}(H)\mathrm{E}(J)^2 \\
= q_1(1-q_1)\frac{\begin{array}{r}12m^2q_2^2 - 24mn_Fq_2^2 + 12n_F^2q_2^2 + 16m_D^2q_2q_3 - 12m_Dmq_2q_3 \\ +24m^2q_2q_3 + 12m_Dn_Fq_2q_3 - 48mn_Fq_2q_3 + 24n_F^2q_2q_3 \\ +7m_D^2q_3^2 - 12m_Dmq_3^2 + 12m^2q_3^2 + 12m_Dn_Fq_3^2 \\ -24mn_Fq_3^2 + 12n_F^2q_3^2\end{array}}{144(q_2+q_3)^2} \\
+ q_1(1-q_1)\left[\frac{2(n_F+m)(q_2+q_3) - m_Dq_3}{4(q_2+q_3)}\right]^2 \\
+ \frac{\begin{array}{r}12m^2q_2^2 - 24mn_Fq_2^2 + 12n_F^2q_2^2 + 16m_D^2q_2q_3 - 12m_Dmq_2q_3 \\ +24m^2q_2q_3 + 12m_Dn_Fq_2q_3 - 48mn_Fq_2q_3 + 24n_F^2q_2q_3 \\ +7m_D^2q_3^2 - 12m_Dmq_3^2 + 12m^2q_3^2 + 12m_Dn_Fq_3^2 \\ -24mn_Fq_3^2 + 12n_F^2q_3^2\end{array}}{144(q_2+q_3)^2}q_1^2
\end{aligned}
$$

Therefore v, which represents the variance of the expectation of $\mathrm{IRE_B}$ given F, can be expressed,

$$
\begin{aligned}
v = \beta^2\Bigg[& q_1(1-q_1)\frac{\begin{array}{r}12m^2q_2^2 - 24mn_Fq_2^2 + 12n_F^2q_2^2 + 16m_D^2q_2q_3 - 12m_Dmq_2q_3 \\ +24m^2q_2q_3 + 12m_Dn_Fq_2q_3 - 48mn_Fq_2q_3 + 24n_F^2q_2q_3 \\ +7m_D^2q_3^2 - 12m_Dmq_3^2 + 12m^2q_3^2 + 12m_Dn_Fq_3^2 \\ -24mn_Fq_3^2 + 12n_F^2q_3^2\end{array}}{144(q_2+q_3)^2} \\
& + q_1(1-q_1)\left[\frac{2(n_F+m)(q_2+q_3) - m_Dq_3}{4(q_2+q_3)}\right]^2 \\
& + \frac{\begin{array}{r}12m^2q_2^2 - 24mn_Fq_2^2 + 12n_F^2q_2^2 + 16m_D^2q_2q_3 - 12m_Dmq_2q_3 \\ +24m^2q_2q_3 + 12m_Dn_Fq_2q_3 - 48mn_Fq_2q_3 + 24n_F^2q_2q_3 \\ +7m_D^2q_3^2 - 12m_Dmq_3^2 + 12m^2q_3^2 + 12m_Dn_Fq_3^2 \\ -24mn_Fq_3^2 + 12n_F^2q_3^2\end{array}}{144(q_2+q_3)^2}q_1^2\Bigg]
\end{aligned}
$$

Step 3) – Observable trial data

The patient sampling, treatmet allocation and mapping of potential outcomes onto observed outcomes were similar to before (see 5.2.2). This time continuous potential dose was mapped onto observed dose as follows,

$$D_{ij} = R_{(B),ij}D_{ij}(R=1) + [1 - R_{(B),ij}]D_{ij}(R=0)$$

Step 4) – Efficacy estimators

Design A: ITT maximum likelihood estimator with random intercept for cluster (therapist) and inclusion of baseline outcome as a covariate. This was the same estimator as used in design option A.

Design B1: Adapted as-treated maximum likelihood estimator of effect of continuous treatment receipt on outcome with random effect for cluster (therapist). The as-treated estimator, without the random effect, was described in Section 4.3.2. The estimator included baseline outcome as a covariate in order to reflect the fact that this was strongly related to outcome at follow-up.

Design B2: Generalised 2SLS estimator with random effect for cluster (therapist). This model, without the random effect, was referred to previously as estimator **E-IV8** (Section 4.5.2). The estimator included baseline outcome as a covariate in both stages in order to reflect the fact that this was strongly related to outcome at follow-up. A random intercept was needed in order to allow for correlations within therapist clusters (this approach was described in Section 4.4.2).

5.3.3 Simulation study design

Seven simulation parameters were varied and three were held constant. I investigated all combinations of these parameters (10,368 trial scenarios) and used 1000 iterations per scenario. The simulation design is generally similar to simulation substudy 1, the main difference being that dose is now continuous. A list of input parameters and their levels is given in Table 5.11.

Those parameters that were varied and were the same as in simulation substudy 1 were sample size, ICC2, cluster size, treatment effect size under design option A (ATE; γ). The parameters that have now changed or been added are treatment effect size under design option B (ACE_{d_1+1,d_0}—ACE_{d_1,d_0}; β), confounding bias due to non-adherence, strength of IV (proportion of latent dose compliers) under design option B, and the strength of compliance within dose compliers (governed by m_D and n_F, which were varied in pairs). The levels that I chose for these parameters were as follows:

- The size of the causal effect in design option B: The target effect (ACE_{d_1+1,d_0}—ACE_{d_1,d_0}; parameter β) represents the change in the causal effect associated with a one-unit

Table 5.11: Summary of input levels of simulation parameters when simulating a continuous measure of treatment receipt.

Parameter	Description	Levels of parameter
Sample size (n)		100; 200; 500; 1000
Treatment effect: β	ACE_{d_1+1,d_0}–ACE_{d_1,d_0} (design option B)	{0.2; 0.5; 0.8} / m
Treatment effect: γ	ATE (design option A)	0.2; 0.5; 0.8
ICC2 (ξ)	Level 2 clustering; due to variance at level of therapist	0.01; 0.02; 0.05; 0.10
Size of clusters (k)		5; 10; 20
Δ	Confounding due to either receipt of treatment in control arm or non-receipt in the intervention arm	{0.2; 0.5; 0.8} / m
Proportion of latent dose compliers (q_1)		0.4; 0.5; 0.6; 0.7; 0.8; 0.9
Proportion of dose never takers (q_2)		0
m	Full dose of treatment	10
Minimum $D(0)$	Minimum dose of treatment received under offer of control for dose always takers and dose compliers	0
m_D (maximum $D(0)$)	Maximum dose of treatment received under offer of control for dose always takers and dose compliers	0; 2; 5; 10 (varied in pairs with n_F)
n_F (minimum F)	Minimum difference in dose between the counterfactual situations for dose compliers	10; 8; 5; 0 (varied in pairs with m_D)
Maximum F	Maximum difference in dose between the counterfactual situations for dose compliers	$10 - D(0)$
α	Effect of outcome at baseline on outcome at follow-up	0.7
ICC1 (ρ)	Level 2 clustering; due to variance at level of catchment area	0.1
Treatment heterogeneity (ϑ)	This is the variance of IRE_A	0.2

Number of trial scenarios: 4 sample sizes, 3 treatment effect sizes, 4 intraclass correlation coefficients (ICC2), 3 cluster sizes, 3 magnitudes of confounding, 6 proportions of dose compliers, and 4 levels of minimum D(0) / maximum F, i.e. 10,368 trial scenarios.

increase in the difference in potential dose between the counterfactual worlds. This was converted into the change in causal effect associated with the maximum difference in potential doses (i.e. full dose). For simplicity, I set $\gamma = m\beta$. The

estimand of interest here is a LATE and was described in Section 4.5.2.

- Confounding bias under design option B is driven by parameter Δ, the confounding effect due either to receipt of treatment in the control arm or to non-compliance in the active intervention arm. This parameter was varied in order to determine small, medium, and large effects of confounding. Note that the parameter represents standardised effects of contamination/non-compliance due to a one-unit increase in contamination or non-compliance. I considered small (d=0.2), moderate (d=0.5) and large (d=0.8) differences, where these differences must be divided by the maximum dose of treatment receipt in order to obtain the change in confounding for a one-unit increase in dose.

- Proportion of dose compliers (q_1) under design option B: this parameter is similar to the proportion of latent compliers in part 1 (p_1) in that it represents the proportion of the population who receive a greater level of active therapy under its offer than they do under offer of control. Proportions were set to the same levels as the proportion of latent compliers in part 1, i.e. q_1 was varied from 0.4 to 0.9 in steps of 0.1.

- The range of possible doses under offer of control under design option B for the dose always takers and dose compliers was given minimum zero and some maximum limit (m_D). This was set to zero, two, five and 10. These maxima enabled an investigation of the effects of the distribution of treatment receipt under offer of control on estimator properties. This parameter was varied in tandem with the following parameter.

- The range of possible differences in dose between offer of control and offer of treatment under design option B for the dose compliers. The upper limit for this is the full dose of treatment minus each participant's level of treatment receipt under offer of control. This subtraction was to ensure that no participant's dose of treatment receipt under offer of treatment was more than m. n_F was set to 10, eight, five and zero. These minima enabled an investigation of the effects of the distribution of the difference in dose between the counterfactual situations on estimator properties.

Parameters m_D and n_F were simulated in four pairs. The four levels of dose compliance that I simulated were given the following labels:

- Full dose compliers. These participants received zero dose when offered control and full dose when offered treatment. This was effectively the same as the generation of binary treatment receipt in part 1 and meant that the dose compliance strata mapped onto the original principal strata.

- Strong partial dose compliers. These participants received a dose of between zero and two when offered control and between eight and 10 when offered treatment.

- Moderate strength partial dose compliers. These participants received a dose of between zero and five when offered control and between five and 10 when offered treatment.

- Weak partial dose compliers. These participants received a dose of between zero and 10 when offered control and between zero and 10 when offered treatment, with a positive difference between offers of treatment and control. This meant that a participant who received a dose of, for example, zero under offer of control would receive a dose of anything between zero and 10 under offer of treatment. A participant who received a dose of, for example, nine under control would receive a dose of between nine and 10 under offer of treatment.

Note that for all four levels of this parameter, all participants in the dose complier stratum received a greater dose when offered treatment compared to when offered control. Also note that as strength of dose compliance becomes weaker, the lower limit of the range of the difference in dose between offer of treatment and contol becomes smaller. This has the effect that with weaker strength of dose compliance, the expectation of the difference in potential dose for the stratum becomes smaller.

Parameters that were fixed and were the same as in simulation substudy 1 were effect of outcome at baseline on outcome at follow-up, ICC1, and treatment heterogeneity. In this simulation substudy I set one further parameter, full dose of active therapy, and also fixed the proportion of dose never takers:

- The full dose of treatment under either offer of control or treatment was m. This was set to 10 on the basis that this represented a plausible upper limit of treatment sessions for a psychological treatment regimen.

- I fixed the proportion of dose never takers (q_2) to zero (i.e. random variable A_{ij} is always one). This simulation substudy was primarily focused on generalising the

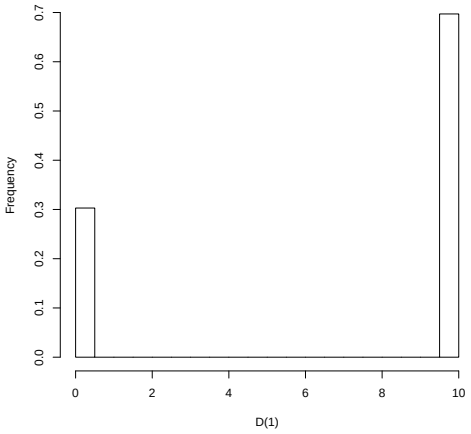
results from substudy 1 to continuous treatment receipt and did not investigate the effect of the presence of dose never takers on the relative efficiency of the design options.

For each trial scenario the same output as described in part 1 was saved.

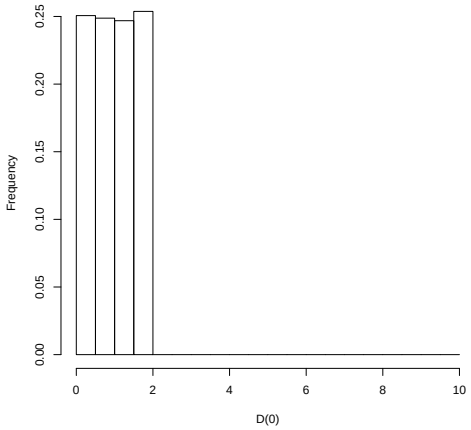
5.3.4 Simulation findings

Distribution of $D(0)$ and F

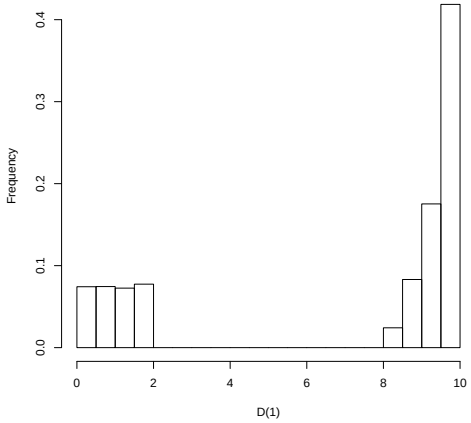
The distributions of dose of active therapy under offer of control and of dose under offer of treatment were investigated under the four different levels of strength of dose compliance. As a reminder, this parameter represented the magnitude of difference in potential dose between trial arms for those participants in the population who would receive a greater dose under offer of treatment compared to control. This was done using data simulations that followed the steps above. I used a large sample size (n=100,000) and a large dose complier stratum of 0.7. For the population with full dose compliers, all participants received no dose under offer of control (hence this was not plotted) and dose under offer of treatment of either zero (30% of sample) or 10 (70% of sample). This is plotted in Figure 5.1a. For the population with strong partial dose compliers, dose under offer of control varied uniformly between zero and two (Figure 5.1b) and dose under offer of treatment was distributed either uniformly between zero and two (30% of sample) or exponentially between eight and 10 (70% of sample; Figure 5.1c). The distribution of this second group was not uniform because for a given participant its upper bound was reduced by the value of dose under offer of control. For the population with moderate strength partial dose compliers, dose under offer of control was a uniform distribution between zero and five (Figure 5.2a) and dose under offer of treatment was distributed either uniformly between zero and five or exponentially between five and 10 (Figure 5.2b). For the population with weak partial dose compliers, the distribution of dose under offer of control was uniform between zero and ten (Figure 5.2c), and dose under offer of treatment ranged between zero and 10 (30% had a difference of zero; Figure 5.2d)). In summary, Figures 5.1b, 5.2a, and 5.2c represent progressive amounts of contamination. Figures 5.1c, 5.2b, and 5.2d represent decreasing dose under offer of treatment (i.e. non-adherence).



(a) Distribution of $D(1)$ when simulating a population with full dose compliers.

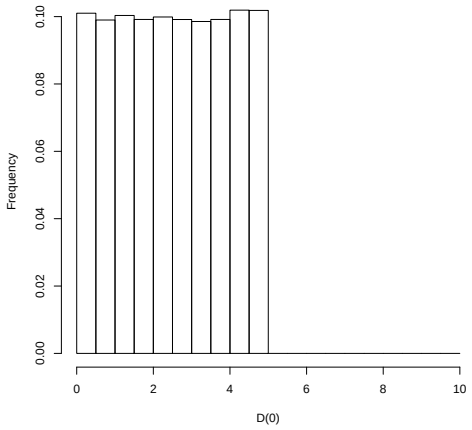


(b) Distribution of $D(0)$ when simulating a population with strong partial dose compliers.

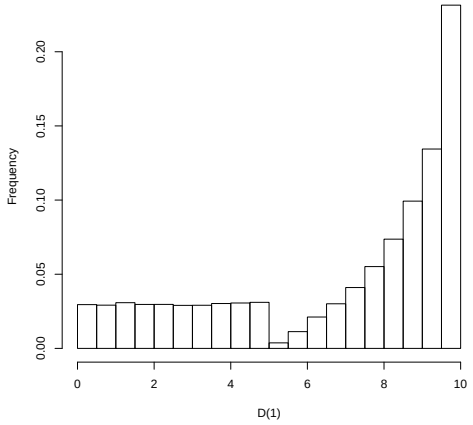


(c) Distribution of $D(1)$ when simulating a population with strong partial dose compliers.

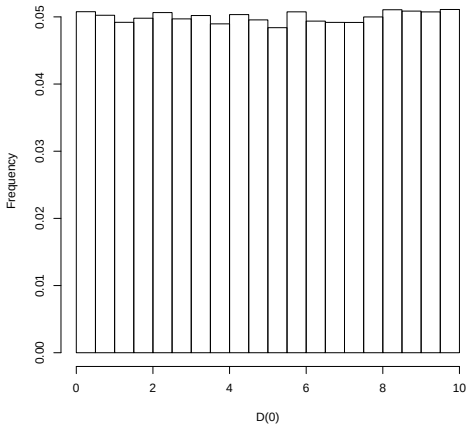
Figure 5.1: Distributions of $D(0)$ and $D(1)$ for the simulations of populations with full dose compliers and strong partial dose compliers. Sample size was set to 100,000, strength of confounding was 0.5, and proportion of dose compliers was 0.7. The distribution of $D(0)$ for the simulations where dose compliers were full compliers is not shown because all participants received a dose of zero under offer of control.



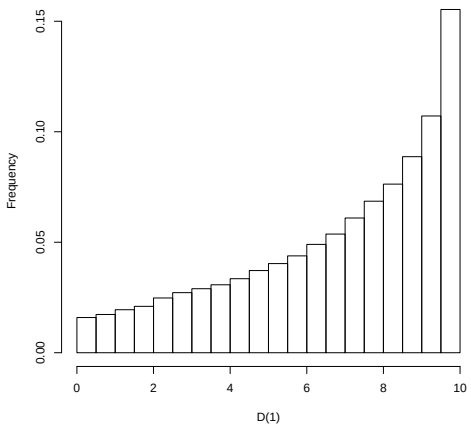
(a) Distribution of $D(0)$ when simulating a population with moderate strength partial dose compliers.



(b) Distribution of $D(1)$ when simulating a population with moderate strength partial dose compliers.



(c) Distribution of $D(0)$ when simulating a population with weak partial dose compliers.



(d) Distribution of $D(1)$ when simulating a population with weak dose compliers.

Figure 5.2: Distributions of $D(0)$ and $D(1)$ when simulating populations with moderate strength and weak partial dose compliers. Sample size was set to 100,000, strength of confounding was 0.5, and proportion of dose compliers was 0.7.

Table 5.12: Absolute bias of as-treated estimator (option B1) rounded to three decimal places, with a continuous measure of treatment receipt and when dose responders were full dose responders. For those with $F > 0$, this assumed that participants received zero dose when offered control and full dose when offered treatment. Results were averaged over ICCs, cluster sizes and magnitudes of treatment effect sizes.

Sample size; strength of confounding	**Proportion of dose responders**					
	0.4	0.5	0.6	0.7	0.8	0.9
100						
0.2	0.145	0.129	0.110	0.087	0.062	0.034
0.5	0.364	0.323	0.272	0.217	0.156	0.083
0.8	0.584	0.517	0.438	0.350	0.250	0.134
200						
0.2	0.146	0.129	0.110	0.087	0.062	0.034
0.5	0.366	0.321	0.275	0.218	0.156	0.084
0.8	0.584	0.516	0.437	0.350	0.248	0.135
500						
0.2	0.146	0.128	0.109	0.087	0.062	0.032
0.5	0.365	0.322	0.274	0.219	0.156	0.084
0.8	0.584	0.516	0.439	0.351	0.250	0.134
1000						
0.2	0.145	0.128	0.110	0.088	0.062	0.033
0.5	0.365	0.322	0.273	0.218	0.156	0.084
0.8	0.585	0.516	0.438	0.350	0.250	0.134

Bias

The results from the data simulations showed a similar story in terms of bias as those for part 1. Specifically, estimator **E-ITT** is the same estimator as part 1, where it was shown to be unbiased (I have not tabularised this again). The as-treated estimator for dose was biased for efficacy under design option B. Table 5.12 shows that there was considerable bias at all levels of sample size, strength of confounding, and size of dose complier stratum. Bias was greatest when strength of confounding was high and size of dose complier stratum was small. Estimator **E-IV8** was an unbiased estimator of $ACE_{d_1+1,d_0} - ACE_{d_1,d_0}$ under design option B. See Table 5.13. The table shows minimal levels of bias at all levels of sample size, strength of confounding, and size of dose complier stratum. The as-treated estimator was dropped and estimators **E-ITT** and **E-IV8** were taken further.

Table 5.13: Absolute bias of IV estimator (option B2) rounded to three decimal places, with a continuous measure of treatment receipt and when dose responders were full compliers. For those with $F > 0$, this assumed that participants received zero dose when offered control and full dose when offered treatment. Results were averaged over ICCs, cluster sizes and magnitudes of treatment effect sizes.

Sample size; strength of confounding	**Proportion of dose responders**					
	0.4	0.5	0.6	0.7	0.8	0.9
100						
0.2	−0.007	−0.001	−0.002	−0.002	0.000	−0.001
0.5	−0.016	−0.009	−0.007	−0.003	−0.002	0.000
0.8	−0.022	−0.011	−0.008	−0.005	−0.002	−0.001
200						
0.2	−0.001	−0.002	0.000	−0.001	0.000	0.000
0.5	−0.005	−0.004	0.001	−0.001	−0.001	0.000
0.8	−0.009	−0.006	−0.005	−0.003	−0.002	0.000
500						
0.2	−0.001	−0.001	0.000	−0.001	0.000	−0.001
0.5	−0.004	−0.001	−0.001	−0.001	−0.001	0.000
0.8	−0.007	−0.002	−0.002	0.000	−0.001	−0.001
1000						
0.2	−0.001	0.000	0.000	0.000	0.000	0.000
0.5	−0.001	−0.001	−0.001	0.000	0.000	0.000
0.8	−0.003	−0.001	−0.001	−0.001	0.000	0.000

Table 5.14: Model SE / true SE of IV estimator (option B2) rounded to three decimal places, with a continuous measure of treatment receipt and when dose responders were full compliers. For those with $F > 0$, this assumed that participants received zero dose when offered control and full dose when offered treatment. Results were averaged over ICCs, cluster sizes and magnitudes of treatment effect sizes.

Sample size; strength of confounding	**Proportion of dose responders**					
	0.4	0.5	0.6	0.7	0.8	0.9
100						
0.2	0.979	0.988	0.983	0.986	0.982	0.978
0.5	0.977	0.980	0.985	0.985	0.988	0.986
0.8	0.975	0.979	0.977	0.987	0.986	0.981
200						
0.2	1.005	1.009	1.000	1.000	0.997	1.001
0.5	0.999	1.002	1.000	1.004	0.996	0.999
0.8	0.997	1.001	1.002	1.001	1.010	0.999
500						
0.2	1.021	1.020	1.019	1.010	1.014	1.016
0.5	1.012	1.019	1.025	1.018	1.015	1.027
0.8	1.021	1.008	1.010	1.016	1.017	1.011
1000						
0.2	1.024	1.027	1.021	1.019	1.024	1.023
0.5	1.026	1.026	1.018	1.020	1.024	1.021
0.8	1.019	1.022	1.015	1.015	1.024	1.025

SE estimation

Summaries of the ratio between model-based SE and true SE for estimator **E-IV8** (Table 5.14) were similar to earlier summaries when treatment receipt was binary. The estimator's model-based SEs increased with sample size; **E-IV8**'s model SEs were closer to the true values at small sample sizes than **E-ITT**'s (see simulation substudy 1).

Monte Carlo standard error

Summaries of the MCSE under design option A (estimator **E-ITT**) and under design option B2 (estimator **E-IV7**) were always less than 0.01. This was similar to simulation substudy 1 – see Tables 5.7 and 5.8.

Relative efficiency

The efficiency of estimator **E-ITT** (design option A) relative to estimator **E-IV8** (design option B2) is summarised by the four types of dose compliers, starting with the full

dose compliers (Table 5.15). The results for the full dose compliers were very similar to those when simulating binary treatment receipt (see Table 5.9). Once again, cells where estimation of efficacy under design option B2 was more precise are coloured in grey. Generally, the results showed that design option B2 was favoured when size of dose complier stratum, cluster size, and ICC were large. The results demonstrated a large range of relative efficiency ratios (0.381-1.997) across the simulated trial scenarios. Design option A was almost always favoured when contamination (proportion of dose always takers) was large, i.e. more than 40%. In these data simulations this level is equivalent to a proportion of dose compliers of 0.6 or less. The picture was mixed at 20% and 30% contamination (design option B2 favoured at high levels of cluster size and ICC). At 10% contamination, design option B2 was favoured at most levels of clustering. At low strengths of clustering (small cluster size and ICC of 0.01 or 0.02) design option A was usually favoured, except in some cases when contamination was very low. As clustering increased, estimation under design option B2 became more efficient. For example, when either cluster size or ICC was of a moderate level, overall design option B2 was favoured in roughly half the scenarios where contamination was 30% or under. At moderate to large cluster sizes (10 or 20) and ICCs (0.05 or 0.1), design option B2 was usually favoured when contamination was 30% or under. Within levels of size of dose complier stratum and ICC, as cluster size increased so did the efficiency ratio. This relationship became more marked as other parameters increased.

Subsequent tables represent generalisations of these results. In these tables the limits on the distributions of dose under offer of control and the difference in potential dose (for the dose compliers) are relaxed. Table 5.16 provides relative efficiency ratios for the simulations of a population containing strong partial dose compliers. These simulations assumed that participants received a dose of between zero and two under offer of control and that dose compliers received a dose under offer of active treatment of between eight and 10. Comparing these results with those for a population with full dose compliers, design option A was favoured at more levels of size of dose complier stratum, cluster size, and ICC. In fact, design option B2 was mainly favoured only when cluster size was 10 or 20, or ICC was 0.05 or 0.1. At these levels and when the proportion of dose always takers was 30% or less, estimation of efficacy under design option B2 tended to be more efficient.

Further generalisations of the distributions of dose under offer of control and dose under

offer of active treatment demonstrated a continuing added advantage for design option A. Table 5.17 shows relative efficiency ratios for a population with moderate strength dose compliers. In this population dose compliers received a dose under offer of control of between zero and five and a dose under offer of active treatment of between five and 10. Design option B2 was only favoured at the highest levels of cluster size and ICC, and when the proportion of dose never takers was 20% or less. Table 5.18 shows the efficiency ratio for a population with weak dose compliers, where both dose under offer of control and dose under offer of active treatment were between zero and 10. The results showed that estimation under design option A was more efficient than under option B2 at every level of size of dose complier stratum, cluster size and ICC.

Table 5.15: Relative efficiency of design options A and B2 for a continuous measure of treatment receipt, where dose responders were full dose responders. For those with $F > 0$, this assumed that participants received zero dose when offered control and full dose when offered treatment. Cells represent ratios of mean estimated standard errors of design option A divided by design option B2. The results are summarised by sample sizes and sizes of dose responder stratum (rows), and intraclass correlation coefficients and cluster sizes (k) (columns). Results were averaged over magnitudes of effect size and strength of confounding.

Sample size;	**ICC=0.01**			**ICC=0.02**			**ICC=0.05**			**ICC=0.1**		
size of dose responder stratum	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$
100												
0.4	0.381	0.413	0.464	0.396	0.441	0.520	0.434	0.532	0.659	0.518	0.633	0.845
0.5	0.490	0.535	0.604	0.516	0.560	0.658	0.564	0.683	0.843	0.646	0.814	1.093
0.6	0.606	0.642	0.719	0.618	0.683	0.797	0.685	0.817	1.012	0.786	0.997	1.307
0.7	0.705	0.772	0.837	0.752	0.791	0.929	0.805	0.947	1.192	0.939	1.185	1.541
0.8	0.815	0.864	0.949	0.864	0.933	1.053	0.925	1.095	1.371	1.080	1.371	1.767
0.9	0.907	0.979	1.076	0.946	1.035	1.204	1.055	1.238	1.511	1.222	1.529	1.997
200												
0.4	0.400	0.435	0.463	0.413	0.460	0.515	0.459	0.536	0.645	0.536	0.666	0.850
0.5	0.516	0.552	0.586	0.535	0.579	0.650	0.586	0.687	0.806	0.682	0.854	1.080
0.6	0.608	0.661	0.698	0.639	0.688	0.780	0.699	0.827	0.994	0.822	1.033	1.286
0.7	0.715	0.764	0.827	0.751	0.824	0.910	0.813	0.970	1.131	0.938	1.203	1.506
0.8	0.821	0.876	0.951	0.853	0.919	1.034	0.938	1.090	1.307	1.090	1.359	1.734
0.9	0.916	0.988	1.044	0.949	1.046	1.162	1.020	1.225	1.502	1.214	1.540	1.963
500												
0.4	0.416	0.441	0.458	0.420	0.463	0.531	0.469	0.553	0.665	0.541	0.692	0.847
0.5	0.518	0.550	0.579	0.534	0.591	0.645	0.589	0.678	0.825	0.686	0.851	1.067
0.6	0.624	0.656	0.693	0.648	0.706	0.777	0.699	0.833	0.987	0.825	1.023	1.313
0.7	0.717	0.753	0.812	0.752	0.828	0.914	0.826	0.960	1.169	0.947	1.196	1.496
0.8	0.818	0.866	0.938	0.854	0.937	1.039	0.938	1.112	1.305	1.082	1.384	1.748
0.9	0.907	0.981	1.042	0.965	1.058	1.179	1.059	1.251	1.504	1.232	1.532	1.976
1000												
0.4	0.407	0.440	0.463	0.428	0.462	0.528	0.467	0.554	0.662	0.536	0.682	0.882
0.5	0.524	0.550	0.587	0.531	0.585	0.653	0.594	0.693	0.823	0.683	0.848	1.098
0.6	0.611	0.643	0.693	0.636	0.689	0.763	0.710	0.826	0.985	0.815	1.024	1.299
0.7	0.720	0.753	0.817	0.756	0.820	0.903	0.809	0.962	1.142	0.950	1.217	1.510
0.8	0.809	0.878	0.947	0.843	0.947	1.047	0.931	1.106	1.335	1.110	1.352	1.751
0.9	0.925	0.964	1.032	0.936	1.059	1.186	1.069	1.218	1.483	1.232	1.534	1.973

Table 5.16: Relative efficiency of design options A and B2 for a continuous measure of treatment receipt, for strong partial dose responders. For those with $F > 0$, this assumed that participants received a dose of between zero and two when offered control and between eight and 10 when offered treatment. Cells represent ratios of mean estimated standard errors of design option A divided by design option B2. The results are summarised by sample sizes and sizes of dose responder stratum (rows), and intraclass correlation coefficients and cluster sizes (k) (columns). Results were averaged over magnitudes of effect size and strength of confounding.

Sample size;	**ICC=0.01**			**ICC=0.02**			**ICC=0.05**			**ICC=0.1**		
size of dose responder stratum	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$
100												
0.4	0.326	0.357	0.398	0.335	0.376	0.438	0.374	0.449	0.555	0.424	0.557	0.707
0.5	0.422	0.450	0.512	0.432	0.482	0.556	0.478	0.571	0.710	0.549	0.700	0.904
0.6	0.518	0.551	0.608	0.521	0.574	0.672	0.583	0.684	0.849	0.658	0.859	1.113
0.7	0.598	0.653	0.720	0.624	0.686	0.790	0.678	0.822	1.004	0.805	1.017	1.317
0.8	0.701	0.751	0.830	0.708	0.787	0.901	0.777	0.931	1.151	0.920	1.164	1.534
0.9	0.776	0.818	0.933	0.786	0.865	1.019	0.891	1.074	1.294	1.027	1.291	1.702
200												
0.4	0.337	0.368	0.388	0.356	0.385	0.436	0.384	0.454	0.553	0.444	0.571	0.728
0.5	0.432	0.460	0.490	0.443	0.472	0.563	0.485	0.579	0.691	0.565	0.708	0.918
0.6	0.530	0.553	0.598	0.529	0.594	0.648	0.597	0.701	0.852	0.685	0.863	1.098
0.7	0.607	0.645	0.691	0.619	0.690	0.770	0.695	0.829	0.986	0.808	1.010	1.272
0.8	0.690	0.753	0.801	0.718	0.779	0.878	0.808	0.927	1.107	0.909	1.167	1.477
0.9	0.779	0.833	0.877	0.796	0.877	0.982	0.878	1.051	1.270	1.030	1.335	1.697
500												
0.4	0.335	0.365	0.384	0.361	0.392	0.432	0.394	0.467	0.556	0.462	0.571	0.721
0.5	0.442	0.465	0.495	0.450	0.494	0.561	0.501	0.584	0.694	0.569	0.737	0.921
0.6	0.519	0.568	0.599	0.542	0.593	0.659	0.592	0.696	0.840	0.701	0.870	1.114
0.7	0.600	0.655	0.697	0.638	0.700	0.767	0.705	0.818	0.985	0.808	1.025	1.283
0.8	0.683	0.744	0.787	0.725	0.795	0.875	0.800	0.930	1.127	0.922	1.166	1.503
0.9	0.770	0.839	0.893	0.800	0.904	0.995	0.899	1.058	1.280	1.030	1.266	1.676
1000												
0.4	0.348	0.367	0.396	0.362	0.400	0.449	0.394	0.469	0.552	0.452	0.573	0.750
0.5	0.436	0.465	0.496	0.446	0.501	0.550	0.502	0.599	0.697	0.576	0.738	0.926
0.6	0.524	0.556	0.594	0.539	0.588	0.679	0.607	0.707	0.839	0.699	0.870	1.117
0.7	0.610	0.642	0.701	0.637	0.692	0.785	0.704	0.814	0.993	0.813	1.027	1.316
0.8	0.686	0.737	0.810	0.715	0.809	0.874	0.789	0.933	1.141	0.917	1.161	1.471
0.9	0.784	0.826	0.878	0.804	0.909	0.972	0.882	1.064	1.260	1.036	1.297	1.679

Table 5.17: Relative efficiency of design options A and B2 for a continuous measure of treatment receipt, for moderate strength partial dose responders. For those with $F > 0$, this assumed that participants received a dose of between zero and five when offered control and between five and 10 when offered treatment. Cells represent ratios of mean estimated standard errors of design option A divided by design option B2. The results are summarised by sample sizes and sizes of dose responder stratum (rows), and intraclass correlation coefficients and cluster sizes (k) (columns). Results were averaged over magnitudes of effect size and strength of confounding.

Sample size;	**ICC=0.01**			**ICC=0.02**			**ICC=0.05**			**ICC=0.1**		
size of dose responder stratum	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$
100												
0.4	0.231	0.255	0.287	0.238	0.268	0.315	0.268	0.324	0.403	0.304	0.402	0.512
0.5	0.305	0.325	0.371	0.312	0.349	0.404	0.345	0.415	0.519	0.399	0.511	0.663
0.6	0.374	0.400	0.442	0.377	0.418	0.491	0.422	0.500	0.621	0.479	0.629	0.819
0.7	0.433	0.475	0.527	0.453	0.500	0.577	0.494	0.601	0.741	0.586	0.746	0.970
0.8	0.510	0.547	0.608	0.516	0.576	0.662	0.567	0.681	0.845	0.676	0.854	1.131
0.9	0.569	0.600	0.684	0.576	0.635	0.748	0.653	0.787	0.950	0.755	0.950	1.254
200												
0.4	0.243	0.266	0.282	0.257	0.280	0.317	0.278	0.330	0.404	0.323	0.417	0.533
0.5	0.312	0.334	0.357	0.320	0.343	0.412	0.353	0.423	0.507	0.412	0.520	0.678
0.6	0.385	0.402	0.437	0.385	0.434	0.475	0.435	0.515	0.625	0.501	0.635	0.808
0.7	0.441	0.470	0.506	0.452	0.504	0.563	0.509	0.609	0.725	0.592	0.742	0.936
0.8	0.503	0.552	0.584	0.524	0.571	0.646	0.592	0.681	0.816	0.666	0.860	1.089
0.9	0.571	0.609	0.643	0.583	0.644	0.723	0.644	0.774	0.934	0.755	0.981	1.254
500												
0.4	0.244	0.266	0.280	0.263	0.286	0.316	0.287	0.342	0.409	0.337	0.420	0.532
0.5	0.321	0.339	0.362	0.328	0.361	0.411	0.365	0.429	0.511	0.415	0.541	0.678
0.6	0.377	0.414	0.437	0.394	0.435	0.484	0.433	0.510	0.618	0.515	0.641	0.821
0.7	0.438	0.478	0.510	0.466	0.512	0.561	0.516	0.601	0.724	0.591	0.755	0.947
0.8	0.498	0.545	0.577	0.529	0.584	0.645	0.585	0.684	0.829	0.675	0.860	1.109
0.9	0.564	0.616	0.657	0.589	0.662	0.731	0.661	0.779	0.945	0.757	0.933	1.236
1000												
0.4	0.254	0.269	0.290	0.264	0.293	0.329	0.287	0.344	0.406	0.331	0.421	0.555
0.5	0.317	0.341	0.362	0.324	0.366	0.404	0.366	0.440	0.513	0.419	0.543	0.684
0.6	0.381	0.405	0.435	0.393	0.430	0.498	0.442	0.519	0.618	0.512	0.641	0.825
0.7	0.445	0.469	0.511	0.465	0.507	0.575	0.515	0.599	0.732	0.597	0.757	0.973
0.8	0.501	0.540	0.595	0.522	0.592	0.642	0.578	0.686	0.841	0.673	0.855	1.085
0.9	0.575	0.606	0.646	0.589	0.668	0.716	0.649	0.784	0.926	0.761	0.954	1.239

Table 5.18: Relative efficiency of design options A and B2 for a continuous measure of treatment receipt, for weak partial dose responders. For those with $F > 0$, this assumed that participants received a dose of between zero and 10 when offered control and between zero and 10 when offered treatment. Cells represent ratios of mean estimated standard errors of design option A divided by design option B2. The results are summarised by sample sizes and sizes of dose responder stratum (rows), and intraclass correlation coefficients and cluster sizes (k) (columns). Results were averaged over magnitudes of effect size and strength of confounding.

Sample size;	**ICC=0.01**			**ICC=0.02**			**ICC=0.05**			**ICC=0.1**		
size of dose responder stratum	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$
100												
0.4	0.004	0.007	0.013	0.007	0.010	0.016	0.008	0.011	0.018	0.015	0.014	0.019
0.5	0.015	0.014	0.025	0.016	0.024	0.022	0.018	0.026	0.035	0.018	0.027	0.047
0.6	0.034	0.043	0.065	0.053	0.047	0.050	0.039	0.074	0.077	0.064	0.068	0.086
0.7	0.093	0.097	0.106	0.074	0.127	0.113	0.115	0.137	0.186	0.141	0.153	0.233
0.8	0.154	0.172	0.201	0.147	0.175	0.217	0.165	0.184	0.282	0.188	0.272	0.374
0.9	0.206	0.216	0.245	0.206	0.203	0.262	0.233	0.280	0.349	0.270	0.336	0.453
200												
0.4	0.016	0.022	0.021	0.023	0.018	0.027	0.021	0.039	0.042	0.028	0.039	0.039
0.5	0.052	0.071	0.075	0.059	0.047	0.105	0.067	0.094	0.102	0.091	0.092	0.151
0.6	0.117	0.127	0.126	0.128	0.144	0.147	0.113	0.168	0.218	0.163	0.209	0.269
0.7	0.155	0.168	0.183	0.161	0.183	0.207	0.184	0.223	0.269	0.212	0.267	0.345
0.8	0.188	0.207	0.217	0.197	0.214	0.244	0.223	0.257	0.308	0.251	0.325	0.415
0.9	0.219	0.231	0.248	0.224	0.247	0.279	0.246	0.299	0.362	0.287	0.373	0.486
500												
0.4	0.083	0.087	0.097	0.091	0.098	0.102	0.090	0.102	0.137	0.110	0.140	0.185
0.5	0.118	0.124	0.135	0.121	0.135	0.153	0.131	0.161	0.192	0.153	0.200	0.252
0.6	0.143	0.156	0.165	0.149	0.166	0.185	0.164	0.195	0.238	0.198	0.247	0.315
0.7	0.169	0.184	0.197	0.180	0.198	0.217	0.200	0.233	0.281	0.229	0.294	0.371
0.8	0.194	0.213	0.224	0.207	0.230	0.254	0.229	0.269	0.325	0.264	0.339	0.434
0.9	0.222	0.242	0.259	0.233	0.260	0.289	0.261	0.308	0.374	0.298	0.368	0.487
1000												
0.4	0.096	0.102	0.110	0.099	0.111	0.125	0.109	0.129	0.153	0.124	0.158	0.211
0.5	0.122	0.132	0.139	0.125	0.141	0.157	0.141	0.171	0.199	0.161	0.211	0.268
0.6	0.149	0.158	0.170	0.153	0.168	0.195	0.172	0.204	0.243	0.201	0.252	0.325
0.7	0.175	0.184	0.200	0.182	0.199	0.226	0.202	0.237	0.290	0.236	0.300	0.386
0.8	0.198	0.214	0.236	0.206	0.234	0.254	0.229	0.272	0.335	0.266	0.341	0.432
0.9	0.228	0.241	0.257	0.234	0.266	0.286	0.259	0.313	0.367	0.303	0.379	0.494

5.4 Discussion

The results showed that, as hypothesised, for a cluster randomised trial design with allocation at level of therapist where clinicians delivered only one therapy, the ITT estimator (accounting for clustered data) provided an unbiased estimate of efficacy (for ATE). This design was named option A. For an individual randomised trial design where therapists delivered both therapies and treatment receipt was measured, the as-treated estimator was biased for efficacy and the IV estimator was unbiased (for CACE). This latter approach was termed design option B2. For low sample sizes, the IV estimator's SE was closer to the truth than the ITT one. For the largest simulated sample sizes the difference was smaller but the IV estimator appeared to slightly overestimate the SE.

For a binary measure of treatment receipt, the relative efficiency of the two approaches was driven by the cluster structure in design option A and the proportion of compliers in the population in design option B2. As all of these parameters increased, efficiency ratios became greater. When there was very little contamination, design option B2 was favoured. Additionally, when the cluster size was moderate or large (10 or 20 in these data simulations) and the ICC was moderate or large (0.05 or 0.1), design option B2 had the advantage when contamination was 30% or less. At the very largest combinations of cluster size and ICC, design option B2 was favoured. When there was non-receipt of treatment in the intervention arm, the results were very similar but the link between size of complier stratum and amount of contamination changed. This non-compliance reduced the size of the complier stratum for estimating efficacy at a given level of contamination. For example, at 20% non-compliance a size of complier stratum of 0.7 equated to 10% contamination. Therefore, at a given level of contamination, the presence of non-compliance effectively reduced the strength of the instrument, thereby reducing the relative efficiency ratio. For a continuous measure of treatment receipt, the difference in potential dose between the counterfactual worlds drove the relative efficiency of the two design options. The greater this difference, the more similar the results were to those for binary treatment receipt. As this difference decreased, efficiency ratios at given levels of cluster size, ICC and size of dose complier stratum tended towards favouring design option A.

The results were consistent with the knowledge that cluster randomisation reduces the efficiency of the estimator of ATE in proportion to the product of the ICC and cluster size

minus one. This is because every extra participant that is recruited in a cluster randomised trial provides less information than that participant would do in an individual randomised trial, due to correlation of outcomes within clusters. The efficiency costs in design option B were the size of the always taker / dose always taker stratum and, for a continuous measure of treatment receipt, weakness of dose compliance (within those participants who would receive greater dose under offer of treatment compared to offer of control). A CACE estimator calculates treatment effect within a subsample of participants who would comply with protocol. As this group of latent compliers becomes smaller, the instrument becomes weaker and the variability of the sampling distribution increases. For continuous treatment receipt, the precision of the estimator of $ACE_{d_1+1,d_0} - ACE_{d_1,d_0}$ is determined by how precisely the gradient of the relationship between the difference in dose between the counterfactual worlds and the size of the treatment effect can be estimated. The relaxation of restrictions on the distributions of dose under offer of control and difference in counterfactual doses had the effect of introducing greater amounts of contamination and non-compliance respectively. This led to less variance in the difference between counterfactual doses and therefore less precision in the magnitude of the relationship between this and treatment effect.

These data simulations have compared two design options with unbiased estimators, held treatment heterogeneity equal between these options, and compared their relative efficiency at plausible levels of a number of parameters. The advice is by no means clear-cut – in some situations cluster randomisation as advocated in the literature is the preferred approach, in other situations it is not. In Chapter 6 the relative efficiency ratios will be plotted in a form that enables the reader to understand how the parameters that drove relative efficiency affected these ratios. The chapter will describe how the results have been made available online. In addition, a trial based on the D6 trial and two hypothetical trials based on typical levels of parameters seen in the other motivating trials and the scoping review will be placed in the context of the results from these simulations.

Chapter 6

Decision support tool

6.1 Background and aims

In Chapter 5 I investigated design approaches for addressing the problem of contamination in RCTs of complex interventions that target efficacy. In this chapter I provide those planning trials with advice regarding optimal design in these RCTs. The following types of trial and contamination process are assumed:

- There are only two trial arms: active arm and control,
- There is no treatment non-compliance in the active arm,
- Contamination is such that contaminated participants either get full active treatment or some dose of it,
- The contamination process is the one most commonly perceived as a problem in mental health research; i.e. therapy in both arms with contamination due to therapists being trained in both interventions and delivering the intervention in the control arm.

I compare the same two design options as described in Chapter 5. These were:

A. Cluster randomisation at the level at which contamination was anticipated (assumed to entirely prevent contamination), with estimation of the ATE (efficacy) and accounting for clustering of outcome data,

B. Individual randomisation, acceptance that contamination will occur, measurement of treatment receipt for all participants, and use of a randomisation-based estimator of the complier average causal effect (efficacy).

Earlier I referred to the second of these approaches as design option B2. I now refer to it as design option B in order to simplify the labels (having dropped the as-treated estimator of design option B1). The scoping review of Chapter 2 showed that design option A is prominent in mental health trials but found no instances of the use of design option B. This could imply that trialists and funders lack necessary information on the merits of the approaches in order to make informed decisions about their use. The results from the simulation studies in Chapter 5 could fill this knowledge gap and thus inform the choice between design options A and B. Therefore I sought to develop a tool that provides investigators with information about the more efficient design option in trials with expected contamination that target efficacy. In particular, I aimed to provide an online interface that makes it easy for the user to judge the relative performance of the two competing options given important planning parameter choices.

To recap Chapter 5, I compared the two design options using Monte Carlo simulations and set parameters to levels that were considered plausible in mental health trials. I completed the simulations in two parts, one with binary treatment receipt and one with continuous treatment receipt. I demonstrated that the estimators under the two trial design options were unbiased and consistent, then assessed their relative efficiency. Therefore the main output from the research was the ratio of estimated treatment effects' standard errors between estimators associated with design options A and B. I was able to identify parameters affecting relative efficiency of the design options from the simulation studies. For binary treatment receipt, these were ICC and cluster size for design option A, and proportion of latent compliers for design option B. When there was no non-compliance, the proportion of latent compliers was simply one minus the proportion of the control arm who received treatment (i.e. one minus the proportion of contaminators). For a continuous measure of treatment receipt, the important parameters were ICC and cluster size for design option A, and the proportion of dose compliers and magnitude of response for design option B. Magnitude of response was the size of the difference in potential doses between treatment offers for those participants who would receive a greater dose under offer of treatment compared to offer of control. Results were summarised in the form of tables, which were stratified by these parameters.

The first aim of this chapter is to demonstrate the efficiency ratios graphically. The second aim is to describe the development of an online application (the "decision support tool") that provides the results from the simulations. The chapter will begin by describing how

the results from Chapter 5 can be demonstrated graphically. This involves assessing the relationship between key design parameters and the relative efficiency of the estimator linked to design option A compared to the one associated with design option B. The chapter will then describe the development of the online decision support tool and provide instructions for its use. Finally, I provide a demonstration of the tool by assessing the better design option for three trials. These include a trial based on the information from the D6 study and two hypothetical trial examples: one with binary treatment receipt and one with continuous treatment receipt. All three target the estimation of treatment efficacy.

6.2 Graphical demonstration of simulation results

Given the finding that, for a binary measure of treatment receipt, there were three parameters that drove the efficiency ratio for a trial of a particular sample size (ICC, cluster size and amount of contamination), I decided to show the relative efficiency of the design options in three dimensions. In the three-dimensional plots, horizontal planes determined the level of clustering (x- and y-axes were ICC and cluster size, respectively) and vertical position was defined by proportion not contaminated in the control arm (i.e. proportion of latent compliers; z-axis). I aimed to plot a surface of equivalence (isosurface) to indicate the plane at which the standard errors for treatment effect under the two design options were the same (i.e. efficiency ratio of one). This was done in R using the 'contour3d' function, which is part of the 'misc3d' package. This function computes and renders isosurfaces using the marching cubes algorithm (Lorensen and Cline, 1987). In essence, this method determines the three-dimensional space which represents some level of an outcome (e.g. level of relative efficiency of one). It does this by assuming a linear change between points and then interpolating where the outcome level of interest is situated. Points are then connected using triangle models of constant density surfaces (e.g. triangular surfaces that represent a level of relative efficiency of one). In totality, these triangles provide the isosurface. As a result, the isosurface maintains connectivity between slices (across the 3D space), surface data, and information on the gradient of the surface. A greater number of three-dimensional points at which the outcome level of interest is known enables the isosurface to be plotted with greater resolution.

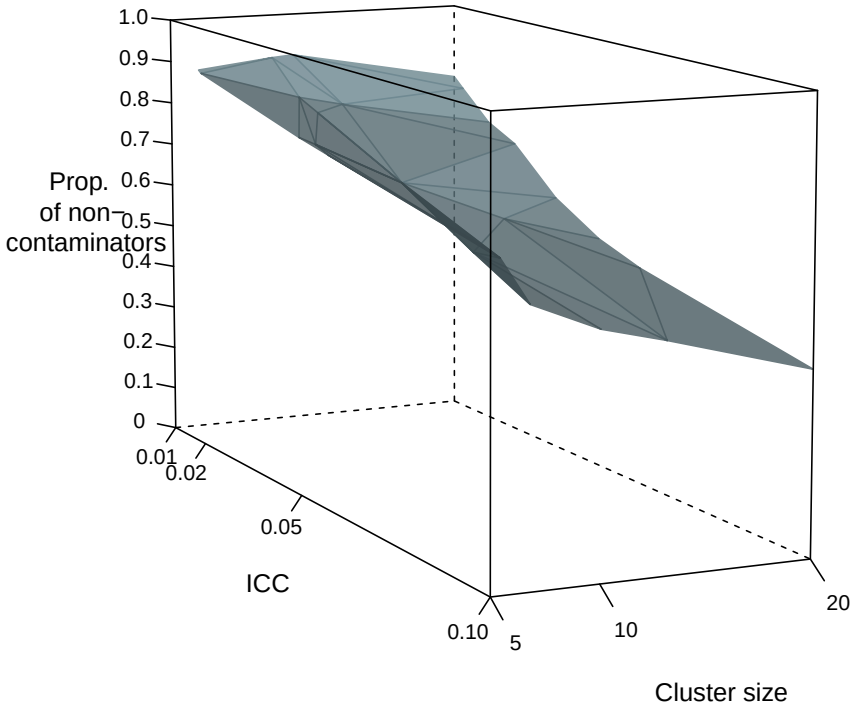
As a demonstration of this, I tried plotting a 3D isosurface for the efficiency ratios that

were calculated for a sample size of 500, treatment effect sizes of 0.5, and strength of confounding of 0.5. This last parameter represented a moderate standardised difference in error terms between complier strata. In order to create this plot, I saved efficiency ratios for the subset defined by the levels of sample size, treatment effect sizes and strength of confounding as a three-dimensional array and then fed this to the 'contour3d' function. I combined the resulting plot of the isosurface with a wireframe which provided axis labels. I drew visible edges of the wireframe in solid black and non-visible edges as dashed lines. The resulting plot is shown in Figure 6.1. The space above the isosurface represents the levels of ICC, cluster size and proportion of non-contaminators at which the standard error for estimation of efficacy is greater for design option A than for design option B. Therefore, above the surface estimation under design option B is more efficient and below the surface estimation under design option A is more efficient. The figure shows that estimation under design option B is more efficient at large proportions of non-contaminators (i.e. little contamination), unless the degree of clustering is very small in which case design option A is favoured (the isosurface hits the top of the box at this level of clustering). As the amount of clustering increases (i.e. greater ICC or cluster size), the space at which design option B is favoured becomes greater. This means that as strength of clustering increases, design option B is favoured at increasing amounts of contamination.

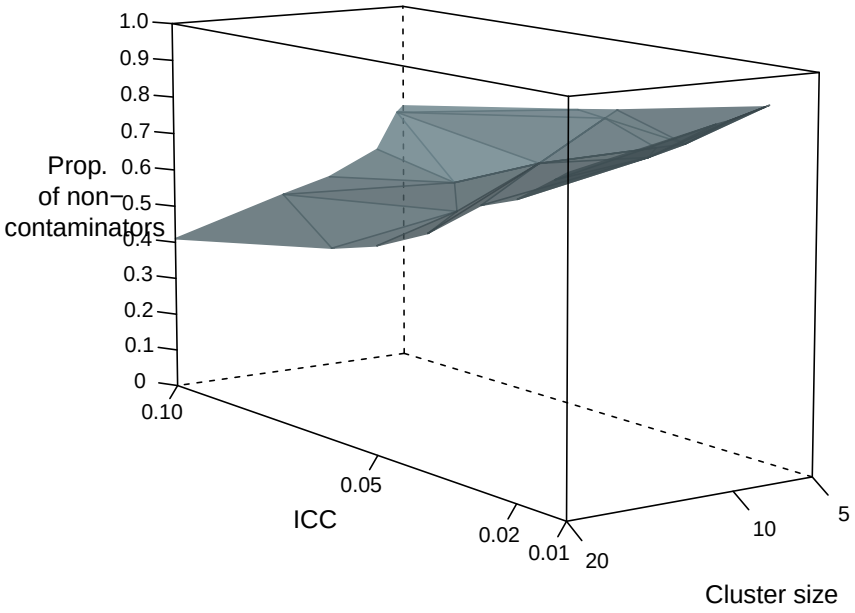
6.3 Online decision support tool

6.3.1 Specification and design

There were three main requirements for the application. Firstly, it should allow the user to enter the levels of the various input parameters that were chosen in the data simulations while providing guidance regarding possible choices for these parameters. This did not include the main parameter for strength of confounding, as this had no bearing on the relative efficiency of the estimators. Secondly, it should provide the mean efficiency ratio (SE under design option A divided by SE under option B) that was found in the data simulations for the level of input parameters that was chosen by the user. Thirdly, the application should provide a plot of the isosurface (for the levels of sample size and treatment effect sizes as chosen by the user) and then plot on the figure the point in three-dimensional space that represented the levels of ICC, cluster size, and proportion of non-contaminators/dose compliers that were selected by the user. This



(a) Three-dimensional plot from frontal perspective of the isosurface representing equivalence of estimator standard errors between design options.



(b) Three-dimensional plot from rear perspective of the isosurface representing equivalence of estimator standard errors between design options.

Figure 6.1: Three-dimensional plot (seen from opposite angles) of the isosurface representing equivalence of estimator standard errors between design options (SE_A / SE_B) for a binary measure of treatment receipt. Sample size was 1000, treatment effect sizes were 0.8, and strength of confounding was 0.5.

would enable the user to place a trial in the context of the 3D surface plot. Finally, the three-dimensional plot should include toggles allowing the user to change the viewing angle of the figure.

As described in the previous paragraph, the input parameters did not include the strength of confounding. This was fixed at its moderate level (i.e. all results shown in the application assumed that strength of confounding was 0.5). In addition it was assumed that the active arm of the trial was not subject to non-compliance in the results shown in the application. This was done to be consistent between the binary and continuous scales on which treatment receipt was measured (non-compliance *per se* was not simulated for the continuous measure of treatment receipt) and because the utility of the application was specifically for deciding how to address the problem of contamination. Therefore, I designed the tool so that the user could enter the levels of sample size, treatment effect sizes, ICC, cluster size, proportion of non-contaminators/dose compliers, and, for a continuous measure of treatment receipt, magnitude of dose compliance.

The application was developed in Shiny, an open source R package for developing online applications. In creating this application I wrote four R scripts that controlled its appearance and functionality:

- "ui.R" script – this is required by Shiny and determines the appearance of the application and the levels of the particular inputs that the user can choose (see Appendix B.1),
- "server.R" script – this is required by Shiny and renders the functionality of the application (e.g. passes the levels of input parameters that were set by the user to the function that plots the 3D figure; see Appendix B.2),
- An R function for printing the efficiency ratio at parameter levels specified by the user (see Appendix B.3),
- An R function for plotting the 3D figure (see Appendix B.4).

6.3.2 Graph toggles

The application included toggles to change the viewing angle of the isosurface plot, alter the zoom, and reset all these controls to their original values. Figure 6.2 demonstrates what the viewing angle toggles do. Arrow (a) represents the rotation in the vertical plane

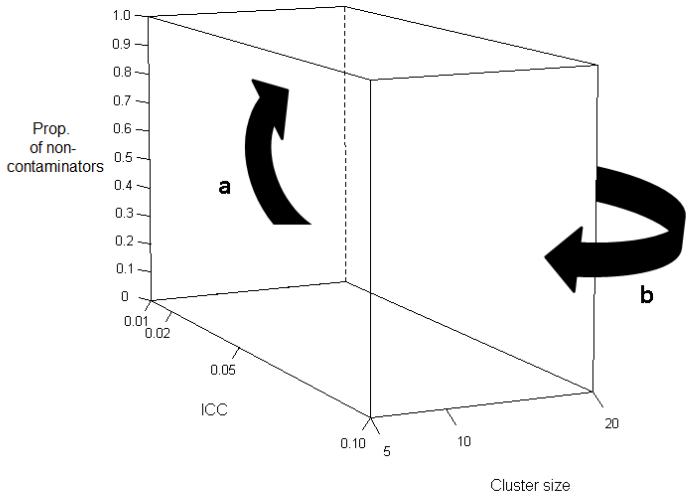


Figure 6.2: Three-dimensional plot of the wireframe box with arrows representing changes in viewing angle for two of the toggles.

(as the toggle is increased the viewing angle moves up) and arrow (b) shows rotation in the horizontal plane (as the toggle is increased the viewing angle moves round to the left).

6.3.3 Online publication

The application has been placed on a server hosted by RStudio and is free to use. It has been published online at: `https://nicholasmagill.shinyapps.io/shiny_app/`. It can be viewed on a screen of any size.

The application has three tabs: user instructions, results (efficiency ratios) for a binary measure treatment receipt, and results for a continuous measure of treatment receipt. The following section provides the user instructions, as given online to users of the tool.

6.3.4 User instructions

This tool compares the efficiency of two design options with consistent estimators of efficacy (the effect of treatment receipt on outcome) in randomised controlled trials where treatment contamination is expected. Contamination is defined as receipt of active intervention within the control arm of a trial. The design options are:

A. Cluster randomisation at the level at which contamination is anticipated to occur with an estimator of ATE that accounts for clustering of data,

B. Individual level randomisation, measurement of treatment receipt, and use of a randomisation-based estimator of efficacy.

Monte Carlo simulations were used to compare these design options. The scenario that is addressed here is a trial with two arms: active intervention and control. It is assumed that there is no treatment non-compliance in the active arm. Participants in the control arm receive either the control treatment, full active treatment, or some dose of active treatment. In design option A, cluster randomisation is assumed to prevent contamination entirely and not to cause bias itself. Any contamination is due to a clinician being trained in both active and control interventions and delivering the active treatment to participants in the control arm.

The tool is presented in two parts. First, with a binary measure of treatment receipt, and second, with a continuous measure of treatment receipt. For a binary measure of treatment receipt, the efficacy estimand is the effect of treatment within a sub-population of participants who would receive treatment when offered it and would receive control when offered it (i.e. the complier average causal effect). For a continuous measure of treatment receipt, the estimand being used is the effect of treatment within a sub-population of participants who would receive the maximum dose of treatment when offered it and no dose when offered control.

The aim of the tool is to provide the ratio of the standard errors of estimates of efficacy under the design options: SE_A / SE_B. A ratio of greater than one would imply that the variance of the estimator of ATE under design option A is greater than that of efficacy under design option B. Or put another way, a ratio of greater than one means that the efficacy estimator of design option B is more precise. The tool provides this ratio at the levels of various parameters, as set by the user. For binary treatment receipt, these parameters are sample size, size of standardised treatment effects, level of ICC, size of clusters, proportion of non-contaminators (this parameter represents the path from random treatment allocation to treatment receipt). For continuous treatment receipt, parameters are sample size, size of treatment effects, ICC, size of clusters, size of dose complier stratum, and the magnitude of the response within this stratum (size of difference between the counterfactual doses). I define the dose compliers as those participants who would receive a greater dose of treatment under offer of active intervention compared to control. The tool also plots a three-dimensional figure of the

surface of equivalence between the two design options (i.e. the plane at which the precisions of the two design options' estimators are equivalent). It is plotted in three dimensions because there are three key variables that drive the relative efficiency of design option A compared to option B: the level of the ICC, cluster size, and proportion of non-contaminator/dose complier stratum.

The user must first choose the sample size options (100, 200, 500 or 1000) and standardised treatment effect sizes (0.2, 0.5 or 0.8). The setting of these parameters enables the generation of the 3D surface plot on the right-hand side of the screen. If the user opts for a binary measure of treatment receipt then he or she needs to determine the ICC (0.01, 0.02, 0.05 or 0.1), cluster size (5, 10 or 20), and proportion of non-contaminators (0.4, 0.5, 0.6, 0.7, 0.8 or 0.9). This proportion is also the proportion of latent compliers, or the population who would receive treatment when offered and would not receive it when offered control. These three parameters are used to plot a coordinate (red ball) in three-dimensional space in the figure. The user may need to use the viewing angle toggles beneath the figure to visualise the position of this point in relation to the surface.

If the user decides to use a continuous measure of treatment receipt, he or she needs to choose the ICC, cluster size, size of dose complier stratum (the proportion of the population who would receive a greater dose under offer of treatment compared to control; 0.4, 0.5, 0.6, 0.7, 0.8 or 0.9), and the magnitude of response within this stratum. This final parameter represents the difference in dose of active treatment under offer of treatment compared to control for the dose compliers. This parameter can be set to one of four levels, which are defined by the minimum dose of active treatment under its offer and the maximum dose of it under offer of control:

- Full dose compliers: participants receive full dose under offer of treatment and nothing under offer of control,
- Strong partial dose compliers: participants receive a dose of between 80% and 100% of maximum dose under offer of treatment and between 0% and 20% under control,
- Moderate strength partial dose compliers: participants receive a dose of between 50% and 100% of maximum dose under offer of treatment and between 0% and 50% under control,

- Weak partial dose compliers: participants receive a dose of between 0% and 100% of maximum dose under offers of treatment and control.

6.4 Demonstration of the decision support tool

I imagine an investigator who is planning a new trial to estimate the efficacy of an experimental therapy compared to a standard therapy. The planning question is how to address the problem of treatment contamination in the design of the trial. Contamination is anticipated if therapists were to be trained in both the new and comparator therapies. It is expected that if this were to happen therapists would provide some control participants with the new therapy or perhaps some dose of it. The design options being considered are those described above (design options A and B). These designs enable unbiased and consistent estimation of efficacy. The online tool can be used to make the decision of which design option is more efficient.

In order the demonstrate the decision making process I use three trial examples. These are one based on D6 and two hypothetical trials: one with binary treatment receipt and one with continuous treatment receipt. These trials constitute examples of existing evidence that can inform parameter choices (treatment effect size, ICC, cluster size, proportion of non-contaminators). For the hypothetical trials, I selected parameter levels that could be considered plausible based on the results of the scoping review in Chapter 2. For the sake of this demonstration, I imagine that all three trial examples aimed to investigate efficacy. I make the same trial assumptions as described earlier. These were that the trial has two arms; there is no treatment non-compliance (no subpopulation of never takers); participants in the control arm receive either full treatment, no treatment, or some dose of treatment; and cluster randomisation prevents contamination and does not cause bias itself.

6.4.1 Planning a trial with a binary measure of treatment receipt

Trial based on D6 study

I imagined an investigator who was planning a new trial that handles the contamination process that D6 was not able to address and targets efficacy. Therefore, D6 represents existing evidence that can be used to inform parameter choices needed as input for the online tool. The design of D6 (according to the trial protocol) anticipated an ICC of 0.05,

cluster size of 15, and standard treatment effect sizes of 0.5. On this basis and assuming 20% drop-out before follow-up along with the loss of two surgeries per arm, the overall sample size requirement was calculated as 432 participants. I entered a sample size of 500, treatment effect sizes of 0.5, ICC of 0.05, and cluster size of 20 into the decision support tool.

The trial's designers did not anticipate the possible level of contamination had it randomly allocated treatment to individuals as opposed to clusters. The tool shows that when the proportion of the control arm who are not contaminated is 60%, the efficiency ratio is slightly greater than one. This output is shown in Figure 6.3. The output shows the efficiency ratio (1.035) and some advice indicating that, because the ratio is more than one, design option B is more efficient. It also provides a plot of the isosurface at the chosen levels of sample size and treatment effect sizes viewed from a side-on perspective. A red ball indicates the three-dimensional space at which the trial is positioned. As a sensitivity analysis, the tool suggests that when the proportion of non-contaminators is 50% or less, design option A is more efficient.

Hypothetical trial with typical levels of relevant parameters

I imagined an investigator who was designing another trial that was targetting efficacy and needed to address contamination in the design. I used the evidence from the scoping review in Chapter 2 in order to choose parameter levels. The review suggested that a typical CRCT was anticipated to have an ICC of 0.05 and cluster size of 10. I entered these parameter levels, an effect size of 0.5, and a sample size of 200 into the decision support tool.

The tool shows that when the proportion of non-contaminators is 80%, the efficiency ratio is more than one. The output from the tool is shown in Figure 6.4. The output shows that the efficiency ratio (1.096) is more than one and indicates that design option B is more efficient. It provides a plot of the isosurface at the chosen levels of sample size and treatment effect sizes. The hypothetical trial is represented by a red ball, which is positioned above the surface. As a sensitivity analysis, the tool provides evidence that when the proportion of non-contaminators is 70% or less, design option A is more efficient.

Subsequent to choosing the design, the investigator should then perform a sample size calculation based on level of power, significance level and treatment effect size. Under

design option A, the sample size calculation would need to inflate for clustering. Under design option B, the calculation would need to target the efficacy estimator.

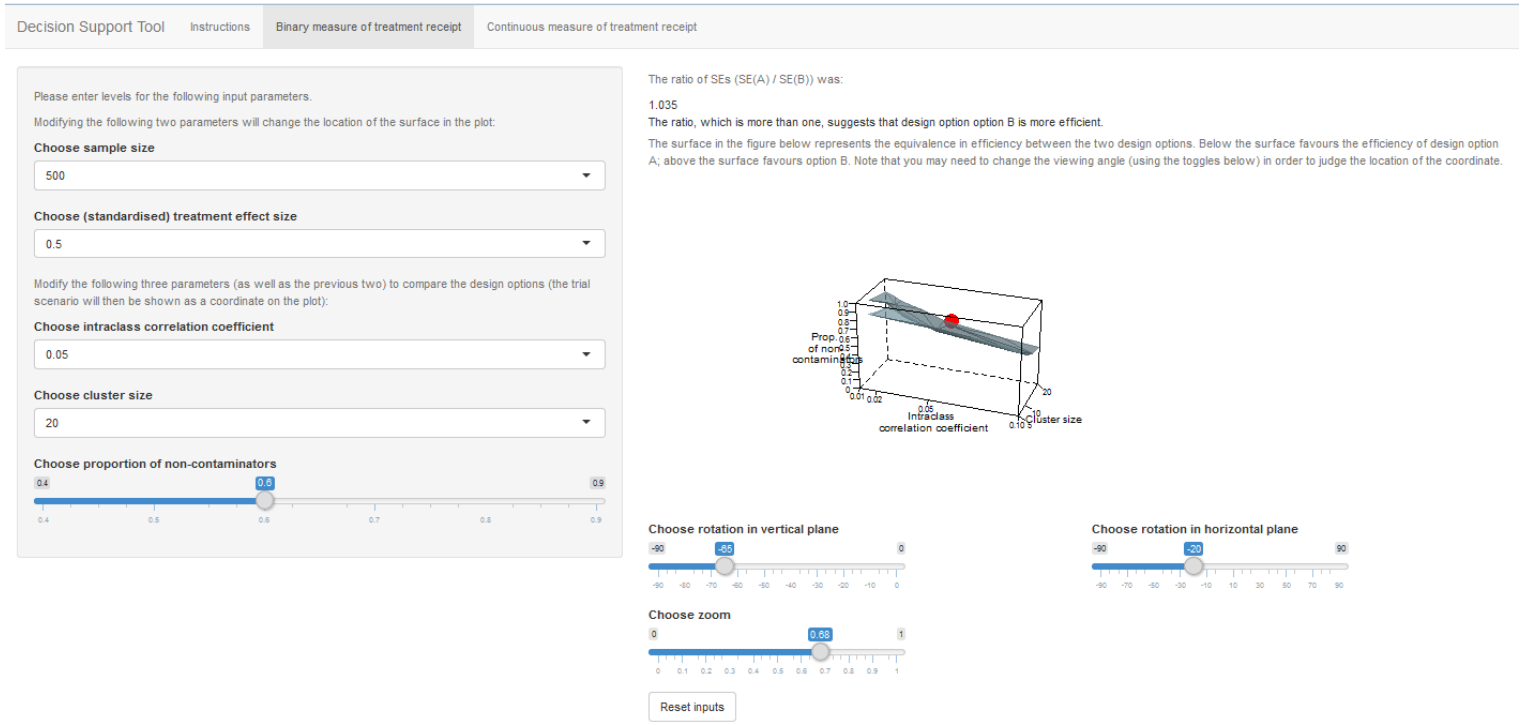


Figure 6.3: Application of the decision support tool to a trial based on the D6 study and defining treatment receipt as binary. At a proportion of non-contaminators of 50%, the ratio of standard errors (under design option A divided by design option B) is more than one and therefore the red ball is above the isosurface. This implies that estimation under design option B would be more efficient.

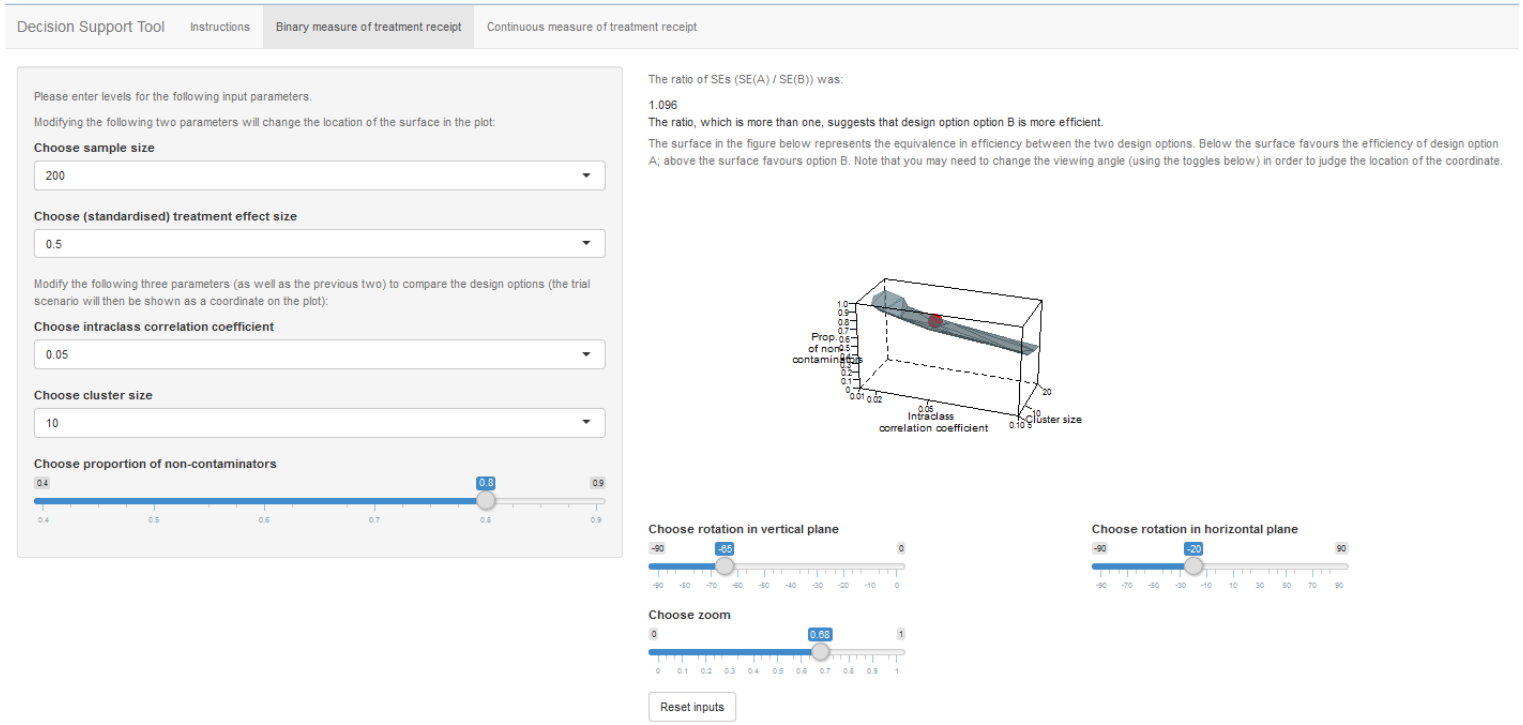


Figure 6.4: Application of the decision support tool to a hypothetical trial with binary treatment receipt and plausible levels of sample size, ICC, cluster size and proportion of non-contaminators. The ratio of standard errors (under design option A divided by design option B) is more than one and therefore the red ball is above the isosurface. This implies that estimation under design option B would be more efficient.

6.4.2 Planning a trial with a continuous measure of treatment receipt (dose)

I imagined an investigator who was planning a trial with a continuous measure of treatment receipt (e.g. number of treatment sessions) and attempting to address contamination in the design. In this case the tool lets the user consider different levels of partial contamination. The parameters that relate to design option A (ICC and cluster size) are the same as in the previous section, but those relating to design option B are different. They are now the size of the dose complier stratum and the magnitude of response within this stratum. As a reminder, this second parameter represented the size of the difference in potential treatment receipt between the offers of treatment and control amongst those participants where this difference would be positive. It also allows for more sources of non-adherence including non-compliance in the active arm. The estimands now being compared are ATE (design option A) and the effect of treatment within a sub-population of participants who would receive the maximum dose of treatment when offered it and no dose when offered control (design option B).

I used the same levels of ICC, cluster size and sample size as described in the previous section. I also used the same expected proportion of dose compliers as the proportion of non-contaminators in the previous section (i.e. 80%). This was with the aim of demonstrating the impact of a change in the magnitude of dose compliance on the efficiency ratio. When the dose compliance parameter was set to "full", the relative efficiency of the design options was similar to that in the previous section. This is because at this level all dose compliers receive full dose under offer of treatment and no dose under offer of control. When I reduced the level of this parameter slightly (i.e. set it to represent "strong partial dose compliers"), the ratio of the standard errors dropped to 0.942. The output from the tool, with parameters set to the levels described, is shown in Figure 6.5. The fact that this ratio was now less than one demonstrates that a small departure from full dose compliance can have a significant impact on the ratio. The reason for this change was that under partial dose compliance there was in effect more non-adherence.

6.5 Conclusion

This chapter has described the plotting of results from Chapter 5 in the form of three-dimensional isosurface figures. For a given sample size and treatment effect sizes, these plots show the space (i.e. levels of ICC, cluster size, and proportion of non-contaminators/dose compliers) at which the two design options are favoured. The chapter has also described the development of an online application that provides the user with information that can be used to decide which of the two competing design options is more efficient under a given research scenario. In particular, it provides the efficiency ratio between design options A and B at a level of the various input parameters that is set by the viewer. The application is reactive (the information that is displayed is dependent on the levels of parameters chosen by the user), free to use, and has been published on the Shiny servers.

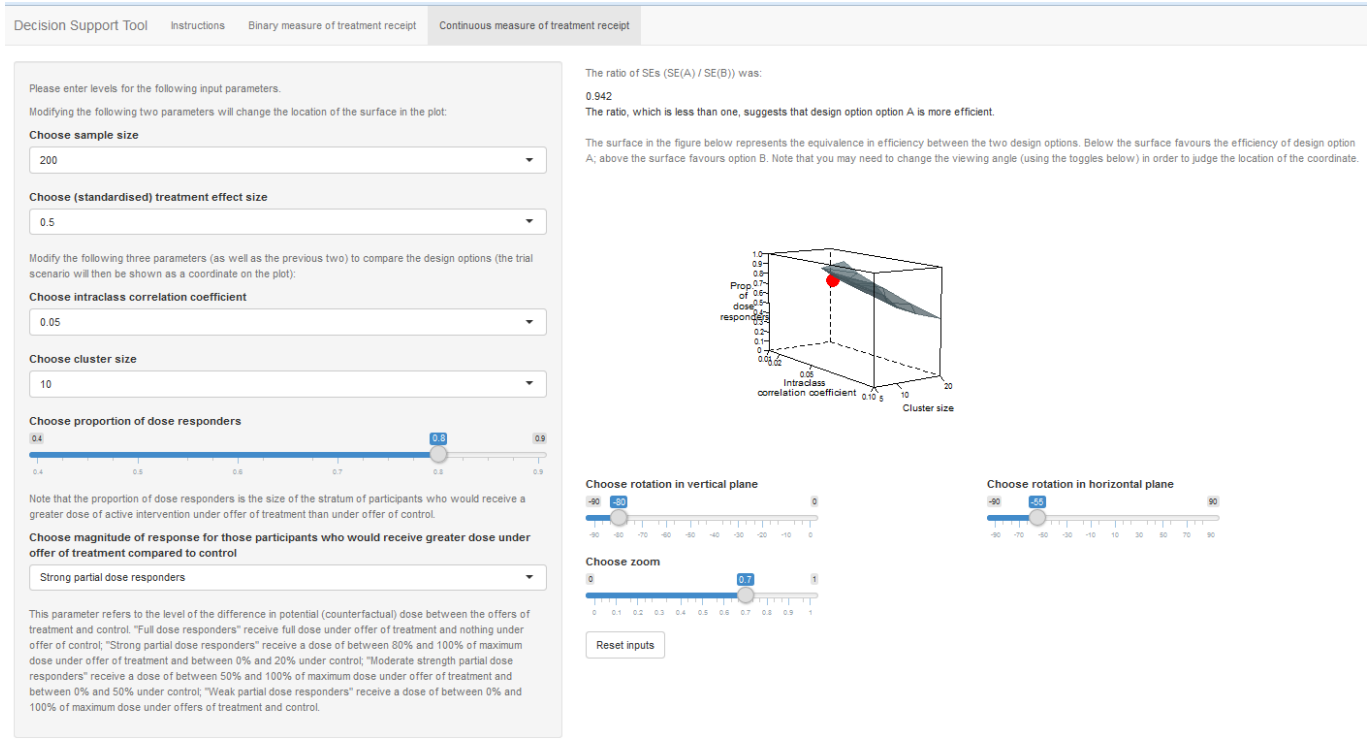


Figure 6.5: Application of the decision support tool to a hypothetical trial with continuous treatment receipt and plausible levels of sample size, ICC, cluster size, proportion of dose compliers and magnitude of dose compliance within this stratum. The ratio of standard errors (under design option A divided by design option B) is less than one and therefore the red ball is below the isosurface. This implies that estimation under design option A would be more efficient.

Chapter 7

Estimating treatment efficacy under treatment contamination: Application to the D6 trial

7.1 Background

I identified asymptotically unbiased estimators of efficacy estimands in Chapter 4. IV and stratification estimators can be used to estimate CACE (binary treatment receipt) and IV methods can be used to estimate $ACE_{d_1+1,d_0}-ACE_{d_1,d_0}$ (continuous treatment receipt) in the presence of non-compliance in the active arm and contamination in the control arm. Previously I detailed the causal assumptions required by these estimators and explained how they can be expanded to accommodate clustered outcome data. In Chapter 5 I demonstrated that design option B (randomisation of individuals, measurement of treatment receipt, and use of randomisation-based efficacy estimator) might be more efficient when contamination is weak as quantified by either a small proportion of the control arm receiving the active intervention (binary treatment receipt) or those in the control arm receiving small doses of active intervention (continuous treatment receipt).

The D6 trial was introduced as a motivating example in Section 1.6. The trial investigated the effect of primary care nurse-led motivational interviewing with CBT skills compared to attention control on HbA_{1c}. The intention was for the experimental treatment to challenge and modify psychological issues and barriers to self-care. The fidelity assessments presented in Chapter 3 showed that non-adherence with randomised intervention (participants allocated to the D6 intervention arm receiving only the control condition

and participants allocated to the control condition receiving some aspect of the attention control intervention) did indeed take place in this trial. Thus contamination occurred despite steps undertaken to prevent it.

This chapter addresses the second part of the secondary research objective. This was to carry out secondary data analyses of the D6 study that use binary or continuous adherence measures to estimate the efficacy of the D6 intervention. Thus I apply design and analysis approach B in practice and compare findings with the original intention-to-treat analysis.

7.2 Aims and hypotheses

For a binary measure of D6 intervention receipt, this chapter aims to answer the question: what is the effect of D6 intervention on primary outcome (glycated haemoglobin, HbA_{1c}) amongst that subpopulation of participants who would receive treatment when offered it and would not receive it when offered attention control? This involves constructing an individual-level measure of adherence (treatment receipt). The aim was to use the MITI and BECCI data that were introduced and described in Chapter 3 to generate a binary exposure variable. This measure will be used to estimate CACE, i.e. the effect of treatment amongst those whose disease management would adhere to protocol (i.e. latent compliers). This will be estimated for HbA_{1c} at all post-randomisation time points (nine, 15 and 18 months after allocation). It was predicted that the estimates of efficacy would be larger than those of effectiveness. It was also predicted that, like the effectiveness estimates that I described in Chapter 1, these would not demonstrate statistical significance.

For a continuous measure of D6 intervention receipt, the questions were as follows. Firstly, what effect is associated with a one-unit increase in dose between the counterfactual situations? Secondly, what is the estimated causal effect for the maximal difference in dose between the counterfactual situations? The aim was to use the treatment fidelity data to construct a continuous measure of individual-level treatment receipt. This was then used to estimate $ACE_{d_1+1,d_0} - ACE_{d_1,d_0}$ for HbA_{1c} at all three post-randomisation time points.

The aim was to use the MITI and BECCI data that were introduced and described in Chapter 3 to generate the endogenous (exposure) variable. The exposure variable was

constructed as either categorical or continuous and then appropriate estimators were applied in order to calculate efficacy for both.

7.3 Method

7.3.1 Statistical issues

As mentioned, the main statistical issue that must be addressed in any efficacy analysis of the D6 trial is the presence of non-adherence. In the following sections I will set out the generation of binary and continuous measures of adherence using the treatment fidelity assessment sample from Chapter 3. I will describe the estimators that were used in the estimation of efficacy (these were introduced in Chapter 4). There were five of these estimators of efficacy (IV and stratification estimators), of which I applied four to the binary adherence measure and one to the continuous adherence measure. I will also explain the methods used for selection of covariates in models.

There were two further statistical issues that were addressed: the presence of missing values in adherence and outcome measures, and the clustering of outcomes due to randomisation of primary care surgeries. The estimators I used assumed that missing adherence values were missing completely at random (MCAR) and made different assumptions regarding missing outcome values. I will explain these assumptions for each model. Whilst the trial was cluster randomised at the level of the primary care surgery, some nurses treated patients in two surgeries which implied the existence of two levels of clustering. Similarly to the primary analysis of the trial, I assumed that nurse clustering would be stronger than surgery clustering (Ismail et al., 2018). I therefore treated the twin GP surgeries as one unit which was equivalent to treating nurse as the primary clustering unit. I will describe how I accounted for this clustering in the analyses using methods that I introduced in Section 4.4.2.

7.3.2 Generation of adherence measures

I considered the MITI and BECCI scales for construction of a binary measure of adherence and only the MITI for quantifying adherence on a continuous scale. I also distinguished adherence with different components of the complex intervention. Specifically, I explored the domains of the MITI, which measured adherence to principles of MI, and the BECCI Practitioner Score (Moyers et al., 2010; Lane et al., 2005), which measured adherence

to the CBT skills part of the intervention. The definitions of these domains are shown in Table 7.1.

Table 7.1: MITI and BECCI domains with definitions.

Domain	Definition
MITI	
Evocation	Extent to which the clinician conveys an understanding that motivation for change and the ability to move towards that change lie mostly with the patient
Collaboration	Extent to which the clinician behaves as if the interview is occurring between two equal partners
Autonomy/Support	Extent to which the clinician supports and actively fosters client perception of choice
Global Spirit	Mean of evocation, collaboration and autonomy / support
Global Empathy	Extent to which the clinician understands or makes an effort to grasp the client's perspective and feelings
Reflection/Question Ratio	Total number of reflections divided by total number of questions
Percent Open Questions	Percentage of total number of questions that were open
Percent Complex Reflections	Percentage of total number of reflections that were complex
Percent MI-Adherent	Percentage of clinician behaviours that were adherent with MI of the total number of adherent and non-adherent behaviours
BECCI	
Practitioner Score	Global measure of skills involved in behaviour change counselling

Binary measure of adherence

I decided to use MITI Global Spirit domain and BECCI Practitioner Score in constructing a binary measure of adherence. This was because MITI Global Spirit is a particularly important domain for the MITI as it represents many aspects of adherence to MI. As described in Table 7.1, it represents the mean of the Evocation, Collaboration and Autonomy/Support subscales (Moyers et al., 2010). Together these make up a large part of the concept of MI. It is measured on a scale of one to five, where one indicates absence of Evocation, Collaboration, and Autonomy/Support. The BECCI Practitioner Score is the mean of the 11 items that constitute the scale. It captures the skill of the clinician in aiming to change the behaviour of a patient who is active and engaged with the treatment (Lane et al., 2005). The D6 trial investigators chose to use this measure to

assess clinicians' ability to provide the CBT skills aspect of the D6 intervention.

I reviewed the literature for a method of using the MITI Global Spirit and BECCI Practitioner Score to define treatment adherence. The authors of the MITI defined thresholds for "Beginning proficiency" and "Competency" for each of the domains (Moyers et al., 2010). For the Global Spirit the "Beginning proficiency" threshold was an average of 3.5 and for "Competency" was an average of 4. These thresholds were based on expert knowledge and were not driven by data. Several randomised controlled trials and pilot studies have used the "Beginning proficiency" threshold as a definition for receipt of treatment (e.g. Simon and Ward, 2014; Larsen et al., 2014; Whittle et al., 2015). I decided to use this lower threshold rather than the upper one because of the nature of the D6 intervention. The treatment was given in conjunction with standard care, meaning that primary care nurses provided patients with advice (e.g. medication adherence) that might be considered as contrasting to the principles of MI.

There are no recommended thresholds for use with the BECCI Practitioner Score. However, the definitions of the five levels of each item are: 0=not at all; 1=minimally; 2=to some extent; 3=a good deal; 4=a great extent. I chose the midpoint between "minimally" and "to some extent" (i.e. 1.5) and defined any scores about this as demonstrating receipt of CBT skills.

In summary, any therapy session recordings that were rated as higher than 3.5 on the MITI Global Spirit domain or greater than 1.5 on the BECCI Practitioner Score were classed as demonstrating adherence to D6 intervention and therefore showed receipt of psychological treatment. Recordings with fidelity ratings that were both lower than these thresholds were classified as showing no treatment receipt. The indicator variable for treatment receipt was coded zero for non-receipt and one for receipt. This was used as the exposure variable in the IV estimators. Similar variables were also created for treatment receipt in each trial arm separately (i.e. observed measures of compliance and contamination) with the difference that these were coded as missing for the trial arm that was not being investigated (e.g. treatment receipt in standard care arm was given missing values for all those allocated to D6 intervention). These were used as the exposure variables in the stratification estimators.

The treatment receipt variable was constructed at the level of the participant, i.e. it represented whether they had been assessed as having received any treatment. Therefore,

a patient was regarded as having received treatment if this was the outcome of at least one audiotape rating.

Continuous measure of adherence

For the continuous measure of treatment receipt, MITI Global Spirit was used as an overall measure of treatment adherence. This was because the main part of the D6 intervention was MI and the main skills of this were included in the measure of Global Spirit. Mean Global Spirit was used as the exposure variable for any participants who had two tape ratings.

7.3.3 Predictors of treatment receipt and missingness

I explored predictors of treatment receipt across both trial arms. Any variables found to be predictive were then included as covariates in the IV models to improve power. The indicator variable for treatment receipt used the definition given in the previous section. Separately, I attempted to identify predictors of the latent classes of always takers and never takers. I did this by searching for predictors of treatment receipt in the control arm (to identify predictors of the always takers class) and then searching for predictors of treatment receipt in the D6 intervention arm (to identify predictors of never takers). Any variables found to be predictive of these classes were then included as covariates in the models for always takers and never takers in the stratification analysis. The aim of this was to increase the ability of the stratification models to identify the latent compliers. I also explored predictors of HbA_{1c} missingness to accommodate observed variables driving missingness in the analyses approaches. Missingness of HbA_{1c} at the primary endpoint (18 months) was coded as zero for missing and one for non-missing.

Pairwise associations between these variables and the main baseline demographic and clinical variables were investigated using logistic regression. The baseline variables available in D6 were gender, marital status, ethnicity, borough, education, age, employment, duration of diabetes and baseline HbA_{1c}. Levels of the categorical variables in this list were coded in the same manner as they were in the primary analysis of the trial.

7.3.4 Efficacy estimators

A list of estimators that were applied to the D6 primary outcome (HbA_{1c}) data is given in Table 7.2 (all of these were introduced in Chapter 4). The first four rows are estimators

that will be applied to a binary measure of treatment receipt; the fifth will be applied to a continuous measure of treatment receipt. Estimators **E-IV7** and **E-IV8** used software that fitted the 2SLS model in one go and therefore used the treatment fidelity assessment sample. For this reason I will summarise this sample in the results section and make comparisons between this sample and the full trial sample. The other estimators, which fitted separate models for treatment receipt and outcome, used all available adherence data and the full trial sample.

Table 7.2: List of estimators from Chapter 4 that were applied to D6 primary outcome data.

Estimator name used in Chapter 4	When introduced	Description
E-IV5	Section 4.5.2, Equation 4.11	Modified Bloom/ratio estimator with bootstrap standard errors, incorporating binary measure of treatment receipt in both trial arms
E-IV6	Section 4.5.2,	Bloom/ratio estimator with binary measure of treatment receipt in both trial arms
E-IV7	Section 4.5.2,	Two-stage least squares estimator with binary measure of treatment receipt in both trial arms
E-STR3	Section 4.5.1	Structural equation mixture model (principal stratification) with binary measure of treatment receipt in both trial arms
E-IV8	Section 4.5.2	Two stage least squares estimator with continuous measure of treatment receipt in both trial arms

7.3.5 Software implementation

HbA_{1c} data were modelled using IV estimators and one stratification estimator (using the principal stratification framework). The IV estimators were performed in Stata 14 (StataCorp, 2015) and stratification models were fitted in MPlus v7.11 (Muthén and Muthén, 2012).

Main analysis of efficacy with binary treatment receipt

I chose the Bloom/ratio estimator (**E-IV6**) for the main analysis of efficacy with binary treatment receipt (i.e. estimation of CACE). As a reminder, the ratio is defined as the estimator of effect of random treatment allocation on outcome (i.e. ITT effect) divided by the estimated proportion of latent compliers. The reasons for choosing this estimator

as the main analysis were twofold. Firstly, it made the least restrictive missing data assumptions. By modelling the effects of random treatment allocation on treatment receipt and outcome separately, the estimators could make full use of the available data and include any baseline predictors of outcome missingness. Secondly, it allowed the inclusion of baseline variables that were predictive of outcome, thereby increasing power. This included all the covariates that were used in the original ITT analysis carried out by the trial team with the exception of time since here analyses were performed separately at each time point. This meant that the variables borough, recruitment phase, and baseline HbA_{1c} were included as covariates. Borough and baseline HbA_{1c} were expected to be predictive of outcome; recruitment was performed in two phases hence its inclusion in the model. The other benefit of this was that it allowed a direct comparison to be made with the effectiveness analysis which included these covariates (i.e. the primary analysis of D6; Ismail et al., 2018).

In order to execute the Bloom/ratio estimator, two regressions were run. Firstly, treatment receipt was regressed on random treatment allocation and the estimated difference in treatment receipt between the trial arms (i.e. proportion of latent compliers) was saved. Then, outcome was regressed on treatment allocation together with any other variables that were predictive of outcome or its missingness and the estimated trial arm difference (the ITT estimate) was saved. A bootstrap program was used to estimate the ratio of these estimates (ITT estimate / proportion of latent compliers) and then compute the standard error for this. The seed was set to 1987, with 200 replications, and resampling at the level of the cluster (nurse). This accounts for clustering of outcome data due to randomisation of primary care surgery / nurse units. Two hundred replications is the upper limit of the range recommended as necessary for normal-approximation confidence intervals (Mooney and Duval, 1993). The code for this model can be found in Appendix C.1. This estimator assumes that missing data are MAR where only observed outcome data and covariates predict missingness. Because the program regressed treatment receipt and outcome separately it made full use of data for each. Therefore the regression model for outcome utilised the full trial sample.

Sensitivity analyses of efficacy with binary treatment receipt

For binary treatment receipt, I applied a number of other estimators in order to assess the sensitivity of the results to inclusion of covariates and assumptions regarding missing outcome data.

I wrote a Stata program to perform the modified Bloom/ratio estimator with treatment receipt allowed to predict missing outcome data and bootstrap standard errors (estimator **E-IV5**). The program regressed treatment receipt on random treatment allocation and saved the estimated intercept (proportion of contaminators) and slope (difference in proportions receiving treatment between the trial arms; also the proportion of latent compliers) as local macros. The observed means of HbA_{1c} stratified by treatment allocation and treatment receipt were also saved. These saved estimates and means were combined to form a ratio as given in Equation 4.11 in Chapter 4. Stata's 'bootstrap' command was used to sample from this ratio in order to obtain standard errors. The same seed, number of replications, and level of resampling (to account for clustered data) were used as before. The Stata code that was used for this is included in Appendix C.2. This estimator assumes that missing data are MAR where observed outcome data and binary adherence predict missingness. It does not allow the inclusion of covariates in the model for the ITT effect. Likewise to estimator **E-IV6**, this estimator modelled treatment receipt and outcome separately, implying that the model for outcome used the full trial sample.

The 2SLS estimator (**E-IV7**) was calculated using Stata's 'ivregress' command. The exposure variable was binary treatment receipt and the instrument was random treatment allocation. Standard errors were constructed using the clustered sandwich estimator, where clusters were nurses, in order to take account of nurse clusters. Model covariates were any variables that were found to be predictive of treatment receipt and HbA_{1c} missingness. Also included were the covariates that were predictive of outcome. These were the covariates included in the ITT analysis by the trial team and were listed in the model for estimator (**E-IV6**). This estimator assumes MAR where observed outcome and the covariates can predict missingness. Because it simulataneously performs both steps of the 2SLS, the estimator utilised only data from the treatment fidelity assessment sample.

The stratification estimator used a mixture model with three latent classes representing the always takers, never takers, and latent compliers. Latent class membership predicted receipt of treatment in the standard care arm (contamination) and receipt of active treatment in the D6 intervention arm (observed compliance). These two parameters were set in such a way that the thresholds enabled perfect prediction by latent class. For example, among the always takers both thresholds were set to the logit value of -15 which meant that members of that class had probability one of receiving treatment under both

allocation to standard care and D6 intervention. The structural model regressed HbA_{1c} on the covariates used in the ITT estimation (borough, randomisation phase, baseline HbA_{1c}) as well as any variables that were identified as predictors of missingness of HbA_{1c} at follow-up. The structural model also regressed membership of the always takers class on any variables that had been found to be predictive of treatment receipt in the standard care arm. Separately it regressed membership of the never takers class on variables found to be predictive of treatment receipt in the D6 intervention arm. This was done to increase precision of the identification of latent compliers (i.e. the remaining class having identified always and never takers) and therefore increase power. The models were fitted under the assumptions of MAR and LI. Under MAR, any baseline variables found to be predictive of missingess were included as covariates in the structual model. Under LI, a dummy variable for outcome response (0=missing, 1=non-missing) was generated and regressed on the same covariates that were in the model for outcome. The compound exclusion restriction was implemented by restricting the parameter for the effect of trial arm on this response variable to be zero for the always takers and never takers. It was not possible to account for clustering of outcome data (due to nurse-patient clusters) in the MPlus mixture model. The code for the model that estimated efficacy at nine months after randomisation under the assumption of latent ignorability can be found in Appendix C.3. The structural part of the SEM includes separate models for latent class membership and outcome. Therefore this estimator used data from the full trial sample.

Continuous treatment receipt

The two-stage least squares estimator for a continuous exposure variable (**E-IV8**) was also generated using Stata's 'ivregress' command. The exposure variable was MITI Global Spirit, where a value of one represented no receipt of treatment and a one-unit increase in dose represented a moderate difference in receipt of MI between the counterfactual worlds. The instrumental variable was random treatment allocation. Covariates included those used in the ITT model (borough, randomisation phase, baseline HbA_{1c}) and any variables that were found to be predictive of missingness at follow-up to facilitate a relaxed MAR assumption. Standard errors were constructed using the clustered sandwich standard error, where clusters were nurses. The software simulataneously performs both steps of the 2SLS and therefore this estimator utilised only data from the treatment fidelity assessment sample.

7.4 Results

7.4.1 Patient sample characteristics

A brief summary of the sample for whom fidelity assessments were available was given in Chapter 3. A fuller description is given here to allow comparison with the trial's target population. Participants were mostly in mid-life, were evenly split between the genders, with no ethnicity being in the majority. About half were married or cohabiting, a large minority had no formal educational qualifications, and most were not in work. Full details are given in Table 7.3. The sample of participants with at least one fidelity assessment did not show any evidence of trial arm imbalance in terms of age, gender, ethnicity, relationship status, duration of illness, and baseline body mass index (BMI). There was some evidence that the D6 intervention group had received a higher level of education and were more likely to be in employment than the standard care group. In addition, there was some suggestion of a relationship between borough and trial arm for this sample.

The sample for whom fidelity assessments were available was similar to the full trial sample in terms of age, gender, relationship status, education level, employment, duration of diabetes (years), baseline HbA_{1c}, and BMI. The fidelity assessment sample comprised a slightly greater proportion of participants of white ethnicity (and lower proportion of African/Caribbean ethnicity), and a smaller proportion of participants from Southwark (and greater proportion from Lewisham) with none from Bexley (in contrast to a small proportion from there in the trial sample).

7.4.2 Adherence with allocated treatments

Using the definition of binary adherence described earlier, 64 (86.5%) participants were deemed to have received treatment in the D6 intervention arm and 38 (49.4%) were regarded as having received it in the standard care arm. Using the terminology of principal stratification, this meant that the estimated proportion of never takers in the population was 13.5% and of always takers was 49.4%. This meant that the estimated proportion of latent compliers was 37.1%.

Unadjusted odds ratios between baseline variables and treatment receipt (across both trial arms) are given in Table 7.4. The table shows that none of the continuous or

Table 7.3: Summary of patient characteristics for the treatment fidelity assessment sample.

Variable	**Level**	**Standard care group (n=77)**	**Intervention (D6) group (n=74)**
Age (years; mean; SD)		59.2 (11.9)	59.7 (10.4)
Gender (n; %)	Male	37 (48.1)	37 (50.0)
	Female	40 (51.9)	37 (50.0)
Ethnicity (n; %)	White	35 (45.5)	33 (44.6)
	African / Caribbean	28 (36.4)	32 (43.2)
	Asian / other	14 (18.2)	9 (12.2)
Relationship status (n; %)	Married / cohabiting	40 (52.6)	36 (48.6)
	Separated / divorced / widowed	18 (23.7)	23 (31.1)
	Single	18 (23.7)	15 (20.3)
Education level (n; %)	A levels or higher	17 (22.7)	22 (30.6)
	O level / GCSE / equivalent	22 (29.3)	30 (41.7)
	No formal qualifications	36 (48.0)	20 (27.8)
Employment (n; %)	In employment	27 (35.1)	33 (44.6)
	Not in employment	50 (64.9)	41 (55.4)
Borough (n; %)	Lambeth	19 (24.7)	40 (54.1)
	Southwark	20 (26.0)	1 (1.4)
	Lewisham	33 (42.9)	10 (13.5)
	Wandsworth	5 (6.5)	23 (31.1)
	Bexley	0 (0)	0 (0)
Duration of diabetes (years; median; IQR)		9 (5-12)	10 (7-13)
HbA_{1c} (mmol/mol; mean; SD)		80.3 (20.9)	79.8 (16.6)
Body mass index (kg / m^2; mean; SD)		31.5 (6.3)	31.4 (5.7)

Table 7.4: Unadjusted odds ratios for possible predictors of treatment receipt.

Variable	Level	Odds ratio	Inference*
Age		0.98	z = -1.27; $p = .21$
Gender	Male		
	Female	0.78	z = -0.70; $p = .48$
Ethnicity	White		
	African / Caribbean	1.43	
	Asian / other	0.44	$\chi^2(2) = 5.35$; $p = .07$
Relationship status	Married / cohabiting		
	Separated / divorced / widowed	0.85	
	Single	1.31	$\chi^2(2) = 0.72$; $p = .70$
Education level	A levels or higher		
	O level / GCSE / equivalent	0.70	
	No formal qualifications	0.34	$\chi^2(2) = 5.89$; $p = .05$
Employment	In employment		
	Not in employment	0.64	z = -1.23; $p = .22$
Borough	Lambeth		
	Southwark	0.27	
	Lewisham	0.31	
	Wandsworth	0.24	
	Bexley	-	$\chi^2(3) = 10.31$; $p = .02$
Duration of diabetes (years)		1.00	$z > -0.01$; $p > 0.99$
HbA_{1c} (mmol/mol)		0.99	z = -1.41; $p = .16$
Body mass index (kg / m^2)		0.96	z = -1.40; $p = .16$

* Tests of the null hypothesis that all odds ratios were equal to one (for each variable).

dichotomous baseline variables predicted treatment receipt. However, tests of the null hypothesis that odds ratios were equal to one were borderline significant for ethnicity, education level, and significant for borough. A test of the null hypothesis for relationship status was not significant.

Investigating predictors of treatment receipt for the two trial arms separately, I found that two baseline variables were weakly predictive of treatment receipt in the control arm. These were gender (odds ratio 0.45, z = -1.70; $p = .09$, 95% CI 0.18, 1.13) and education (test of null hypothesis that odds ratios were equal to one, $\chi^2(2) = 5.71$, $p = .06$). Two variables were predictive of treatment receipt in the D6 intervention arm. These were both strongly predictive and were borough (test of null hypothesis that odds ratios were equal to one, $\chi^2(1) = 8.59$, $p < .01$) and baseline BMI (odds ratio 1.22, $z = 2.80$; $p = .01$, 95% CI 1.06, 1.41).

7.4.3 Missingness of HbA_{1c} and predictors of it

Considering the full trial sample HbA_{1c} data at all time points, the proportion of missingness ranged between 33-58%. The proportion of missing data was greatest at 15 months after randomisation and lowest at 18 months (this was the trial's primary endpoint). It was slightly higher in the standard care group than intervention group at 15 months; the proportions were similar at nine and 18 months. This description was similar to the proportion of HbA_{1c} missingness in the treatment fidelity assessment sample, with the caveat that overall missingness was a few percentage points lower in the fidelity analysis dataset.

Unadjusted odds ratios between baseline variables and HbA_{1c} response (i.e. non-missingness) for the full trial sample at 18 months after randomisation are given in Table 7.5. The table shows that none of the continuous or dichotomous baseline variables predicted HbA_{1c} response, with the exception of baseline HbA_{1c} which was borderline significant. A test of the null hypothesis that odds ratios were equal to one was highly significant for ethnicity. Tests of the null hypotheses were not significant for relationship status, education, and borough.

There was no evidence that treatment receipt predicted missingness of HbA_{1c} at 18 months after randomisation. Amongst those participants who did not receive treatment, 35 (71.4%) had non-missing outcome data whilst this figure was 72 (70.6%) amongst those who did receive treatment.

7.4.4 Effectiveness assessment

HbA_{1c} for both the full trial and the treatment fidelity assessment samples is summarised by trial arm and time point in Table 7.6. Means and 95% confidence intervals for the full trial sample are plotted in Figure 7.1. For the full trial sample, the first part of the table and the figure show that mean HbA_{1c} decreased in both trial arms over time during the first 15 months of follow-up. The level of HbA_{1c} increased slightly between months 15 and 18. At all time points the means were similar between the trial arms.

Estimates of effectiveness using an ITT analysis for the full trial sample are given in Table 7.7. There was no evidence for a difference in HbA_{1c} between trial arms at any time point. This raised the question of whether an estimator of efficacy instead of effectiveness might show a larger effect.

Table 7.5: Unadjusted odds ratios for possible predictors of HbA_{1c} response at 18 months after randomisation.

Variable	Level	Odds ratio	Inference*
Age		1.00	$z = -0.05$; $p = .96$
Gender	Male		
	Female	1.17	$z = 0.66$; $p = .51$
Ethnicity	White		
	African / Caribbean	2.22	
	Asian / other	3.09	$\chi^2(2) = 14.47$; $p < 0.001$
Relationship status	Married / cohabiting		
	Separated / divorced / widowed	0.85	
	Single	1.25	$\chi^2(2) = 1.22$; $p = .54$
Education level	A levels or higher		
	O level / GCSE / equivalent	0.91	
	No formal qualifications	0.95	$\chi^2(2) = 0.09$; $p = .96$
Employment	In employment		
	Not in employment	0.95	$z = -0.20$; $p = .84$
Borough	Lambeth		
	Southwark	0.92	
	Lewisham	0.71	
	Wandsworth	0.83	
	Bexley	0.22	$\chi^2(4) = 6.20$; $p = .18$
Duration of diabetes (years)		1.01	$z = 0.69$; $p = .49$
HbA_{1c} (mmol/mol)		0.99	$z = -1.93$; $p = .05$
Body mass index (kg / m^2)		1.02	$z = 0.84$; $p = .40$

* Tests of the null hypothesis that all odds ratios were equal to one (for each variable).

Table 7.6: Summary of HbA_{1c} (mmol/mol) at outcome time points for both the full trial and treatment fidelity assessment samples.

HbA_{1c} (mmol/mol)	**Standard care group (n=170)**	**Intervention (D6) group (n=164)**
Whole trial sample		
Baseline (mean; SD)	80.08 (19.12) n=168	80.98 (17.07) n=164
Nine months after randomisation (mean; SD)	75.94 (20.15) n=108	76.38 (18.91) n=101
Fifteen months after randomisation (mean; SD)	72.74 (19.22) n=71	73.12 (15.81) n=80
Eighteen months after randomisation (mean; SD)	73.81 (19.89) n=109	74.11 (18.74) n=110
Treatment fidelity assessment sample		
Baseline (mean; SD)	80.31 (20.88) n=77	79.79 (16.64) n=74
Nine months after randomisation (mean; SD)	72.70 (17.60) n=53	76.83 (18.59) n=46
Fifteen months after randomisation (mean; SD)	69.81 (17.46) n=37	74.12 (15.79) n=43
Eighteen months after randomisation (mean; SD)	70.70 (16.28) n=54	69.89 (15.79) n=53

Table 7.7: Estimated difference in HbA_{1c} (mmol/mol) between treatment groups at outcome time points using effectiveness estimator (ITT analysis).

Time since randomisation	**Estimated difference (mmol/mol)**	**Standard error**	***z*-score and *p*-value**	**95% confidence interval**
Whole trial sample				
Nine months after randomisation	1.00	2.00	$z = 0.50$; $p = .62$	-2.91, 4.92
Fifteen months after randomisation	1.81	2.23	$z = 0.81$; $p = .42$	-2.57, 6.18
Eighteen months after randomisation	-0.15	2.60	$z =$ -0.06; $p = .96$	-5.24, 4.95
Treatment fidelity assessment sample				
Nine months after randomisation	7.15	3.53	$z = 2.02$; $p = .04$	0.23, 14.08
Fifteen months after randomisation	5.16	3.10	$z = 1.67$; $p = .10$	-0.91, 11.23
Eighteen months after randomisation	0.98	2.77	$z = 0.35$; $p = .72$	-4.45, 6.41

Estimates were calculated separately at the three post-randomisation time points. Covariates were the same as those included in the primary analysis: trial arm, borough, recruitment phase, baseline HbA_{1c}. Clustering of outcome data was accounted for using the clustered sandwich estimator.

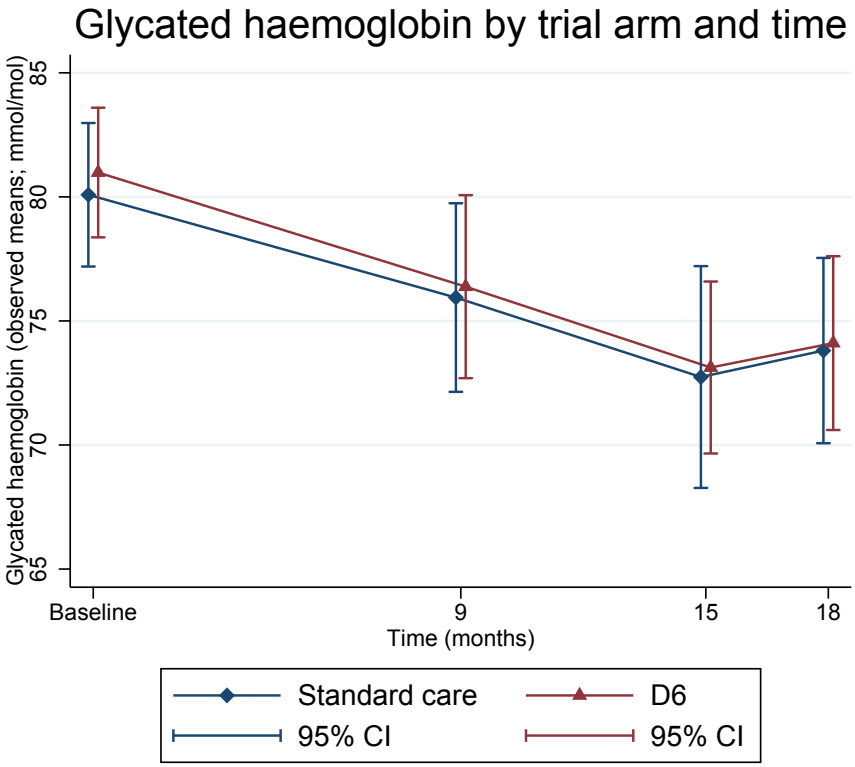


Figure 7.1: Plot of means and 95% confidence intervals of standard care and D6 intervention arms over time for the primary analysis (full trial) sample.

The trends in HbA_{1c} for only those participants who had a record of treatment fidelity were a little different to those for the full trial sample (Table 7.6). There were approximately half as many data points in the fidelity analysis sample as the full trial sample. Baseline HbA_{1c} levels were similar between the samples. However, HbA_{1c} declined more steeply over the first nine months in the standard care group in the fidelity analysis sample compared to the larger sample. Therefore, at nine months after randomisation there appeared to be a small difference in HbA_{1c} levels between the trial arms, with the standard care group demonstrating better control of HbA_{1c}. At 15 months this gap was maintained but at 18 months the difference had disappeared.

For the treatment fidelity assessment sample, the inferential ITT analysis demonstrated weak evidence of a difference between the trial arms at nine months after randomisation (Table 7.7). The estimate was positive implying that HbA_{1c} was greater (i.e. worse) in the D6 intervention arm. This difference diminished somewhat at 15 months and particularly at 18 months; the differences were not significant at both of these time points. This meant that there were some differences in the estimates at nine and 15 months between the full trial and treatment fidelity assessment samples. The difference was small at 18 months.

7.4.5 Efficacy assessment

Efficacy estimates at the three post-randomisation time points for each of the estimators where treatment receipt was defined on a binary scale are given in Table 7.8. What follows is a description of the efficacy analyses using a binary measure of treatment receipt. I have split this into the main analysis using estimator **E-IV6** followed by sensivity analyses using the other estimators proposed in Chapter 4 (estimators **E-IV5**, **E-IV7**, **E-STR3**). Finally, I describe the efficacy analysis with continuous treatment receipt using estimator **E-IV8**.

Main analysis of efficacy with binary treatment receipt variable

Estimator E-IV6 - Bloom/ratio estimator with bootstrap standard error

Estimates of CACE using estimator **E-IV6** were positive (in direction of higher/worse HbA_{1c} in the D6 intervention arm) at all three time points. The numerator part of the ratio, which was the regression of outcome on random treatment allocation, covaried for the same baseline variables that were used in the ITT analysis together with variables

Table 7.8: Estimated differences in HbA_{1c} (mmol/mol) between treatment groups at outcome time points using efficacy estimators with binary measure of treatment receipt. These are estimates of CACE.

Estimator name	Time since randomisation	Estimated difference (mmol/mol)	Standard error	*z*-score and *p*-value	95% confidence interval
Main analysis of efficacy					
E-IV6*	9 months	3.73	12.65	$z = 0.30$, $p = .77$	-21.06, 28.53
	15 months	5.77	9.67	$z = 0.60$, $p = .55$	-13.19, 24.73
	18 months	1.19	11.67	$z = 0.10$, $p = .92$	-21.69, 24.07
Sensitivity analyses of efficacy					
E-IV5	9 months	8.92	13.43	$z = 0.66$, $p = .51$	-17.40, 35.24
	15 months	9.68	17.21	$z = 0.56$, $p = .57$	-24.04, 43.40
	18 months	-1.87	9.01	$z =$ -0.21, $p = .84$	-19.52, 15.79
E-IV7*	9 months	25.84	11.11	$z = 2.33$, $p = .02$	4.07, 47.62
	15 months	12.45	8.13	$z = 1.53$, $p = .13$	-3.49, 28.39
	18 months	5.24	6.00	$z = 0.87$, $p = .38$	-6.51, 17.00
E-STR3† (assuming MAR)	9 months	0.63	6.41	$z = 0.10$, $p = .92$	-11.94, 13.20
	15 months	2.16	6.58	$z = 0.33$, $p = .74$	-10.73, 15.05
	18 months	7.88	6.04	$z = 1.30$, $p = .19$	-3.96, 19.72
E-STR3† (assuming LI)	9 months	-4.78	13.95	$z =$ -0.34, $p = .73$	-32.12, 22.55
	15 months	1.03	7.12	$z = 0.14$, $p = .89$	-12.92, 14.97
	18 months	7.14	8.07	$z = 0.89$, $p = .38$	-8.68, 22.96

* Covariates in the model for outcome included borough, recruitment phase, baseline HbA_{1c}, ethnicity and education.

† Covariates in the model for outcome included borough, recruitment phase, baseline HbA_{1c} and ethnicity.

that were found to be predictive of observed treatment receipt. These estimates should be compared to the effectiveness estimates for the full trial sample (see first part of Table 7.7). The efficacy estimates showed a similar pattern to the effectiveness estimates (largest at 15 months and smallest at 18 months) and were larger at all time points. The precision of the efficacy estimates was considerably less than that of the effectiveness estimates.

Sensitivity analyses of efficacy

Estimator E-IV5 - modified Bloom/ratio estimator (incorporating observed adherence) with bootstrap standard errors

Estimates of CACE using estimator **E-IV5** were positive (in direction of higher or worse HbA_{1c} in the D6 intervention arm) at nine and 15 months and negative at 18 months. Likewise to efficacy estimator **E-IV6**, these estimates should be compared to the first set of effectiveness estimates (the full trial sample). Efficacy estimates at all three time points were in the same direction as the effectiveness estimates and standard errors were considerably larger. It should be noted that these estimators did not covary for baseline variables that were predictive of outcome (and featured as covariates in the effectiveness estimators) or for baseline variables that were predictive of outcome missingness. In addition, the ratio estimator of efficacy allowed observed adherence to predict missingness of outcome which the ITT analysis did not.

Estimator E-IV7 - two-stage least squares estimator

Estimates of CACE using estimator **E-IV7** were positive (in direction of worse HbA_{1c} in D6 intervention arm) at all time points. Estimates were considerably larger than estimates for the previous IV estimators of CACE. Estimates should be compared to the effectiveness estimates for the treatment fidelity assessment sample. The efficacy estimates were in the same direction and showed the same pattern, i.e. largest at nine months (when it was statistically significant) and diminishing after this. Precision was substantially lower at all time points.

Estimator E-STR3 - stratification estimator with structural models for outcome, always takers class, and never takers class

Estimates of CACE using estimator **E-STR3** were positive (in direction of worse HbA_{1c} in D6 intervention arm) at all time points under the assumption of MAR. The smallest estimated effect was at nine months and largest at 18 months. Under the assumption of

LI, the results at 15 and 18 months were similar to those under the assumption of MAR, but the estimate at nine months was in the opposite direction and slightly larger. For the MAR models, the estimated proportions of latent compliers ranged between 37.1% and 42.3% over the three time points; for the LI models, this range was 26.4% to 41.0%. The stratification estimates under the assumption of MAR can be roughly compared to the effectiveness estimates for the full trial sample. The stratification efficacy estimates were larger in magnitude than the effectiveness estimates but patterns across time points were different. This may be because the models are not entirely comparable as the stratification estimators use different covariates in the structural model for outcome (they include predictors of class membership). Standard errors were larger than those of the effectiveness estimates, particularly at nine months under the LI assumption.

Analysis of efficacy with continuous treatment receipt variable

Estimator E-IV8 - two-stage least squares estimator

The estimates of efficacy using estimator **E-IV8** are shown in Table 7.9. They estimate the causal treatment effect on HbA_{1c} associated with a one-unit increase in the difference of MITI Global Spirit between the counterfactual situations. This is to say that they estimate ACE_{d_1+1,d_0}–ACE_{d_1,d_0}. The pattern of estimates can be compared to the pattern of effectiveness estimates for the treatment fidelity assessment sample (i.e. magnitude of effect was greatest at nine months then reduced). At 18 months, the results imply that the predicted effect of treatment at the maximal difference in dose between the counterfactual situations was 5.67 mmol/mol (SE 6.79, $z = 0.83$, $p = .40$, 95% CI -7.64, 18.99). Estimated causal treatment effects (and 95% confidence intervals) for the maximum difference in dose between the counterfactual worlds at all three post-randomisation time points are shown in Figure 7.2.

7.5 Discussion

In this section I will summarise the main results from the efficacy analyses. I will then return to the statistical issues that I identified at the beginning of the methods section of this chapter. For binary treatment receipt, I will explore the impacts of inclusion of covariates in models and assumptions regarding missing data. I will interpret the estimates of efficacy for a continuous measure of treatment receipt and explore the estimator's assumptions. Finally, I will draw some conclusions on the D6 trial and

Table 7.9: Estimated differences in HbA_{1c} (mmol/mol) at outcome time points using efficacy estimators with continuous measure of treatment receipt. These are estimates of ACE_{d_1+1,d_0}–ACE_{d_1,d_0}.

Estimator name	Time since randomisation	Estimated difference (mmol/mol)	Standard error	*z*-score and *p*-value	95% confidence interval
E-IV8*	9 months	5.27	2.12	$z = 2.49$, $p = .01$	1.13, 9.42
	15 months	2.89	2.04	$z = 1.42$, $p = .16$	-1.11, 6.89
	18 months	1.42	1.70	$z = 0.83$, $p = .40$	-1.91, 4.75

* Covariates in the model for outcome included borough, recruitment phase, baseline HbA_{1c}, ethnicity and education.

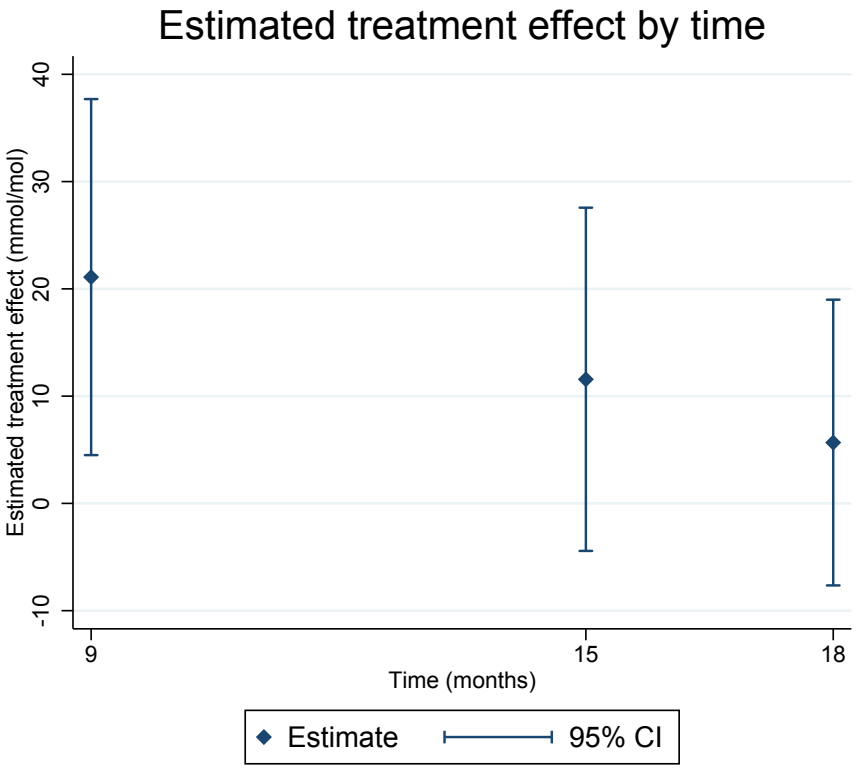


Figure 7.2: Plot of estimated causal treatment effects associated with maximal difference in MITI Global Spirit between the counterfactual situations at all post-randomisation time points.

intervention.

Overall, the efficacy estimators provided little evidence of a treatment effect and no evidence of treatment benefit in the D6 trial. When treatment receipt was binary, the main efficacy estimates at the three post-randomisation time points were positive, meaning that the level of HbA_{1c} in the experimental treatment group was estimated to be slightly greater, and therefore worse, than in the standard care group. Point estimates of efficacy and standard errors were larger than those of effectiveness at the same time points. The main efficacy analysis showed that estimates peaked at 15 months. This pattern was consistent with the estimates of treatment effectiveness. These findings were entirely consistent with my predictions at the beginning of this chapter. When treatment receipt was continuous, the estimated effects were also positive at all time points. The largest estimated treatment effect was at nine months after randomisation (and was statistically significant); after this, results diminished with time. The pattern of estimates was similar to an effectiveness analysis that was restricted to those participants who made up the treatment fidelity assessment sample. These findings were roughly consistent with my predictions, although the significant result at nine months after randomisation was a surprise. I will consider this further below.

The main statistical issues were how to estimate efficacy in the presence of non-adherence to randomised treatment allocation, the presence of missing adherence and outcome measures, and clustering due to randomisation of primary care sugeries. All estimators assumed that missing adherence data were MCAR, or equivalently it was assumed that the non-missing data constituted a random subsample of the full trial sample. I addressed clustering by either bootstrap sampling at the level of the cluster (for the ratio estimators) or using the clustered sandwich estimator (for the 2SLS models). The estimators that I applied to D6 differed in whether or not they could accommodate baseline predictors of outcome. Estimators **E-IV6**, **E-IV7**, **E-STR3** and **E-IV8** allowed inclusion of such covariates in the model for outcome. They also differed in the strictness of their missing outcome data assumptions. Certain estimators could accommodate outcome data from participants whose adherence measures were missing (i.e. **E-IV5**, **E-IV6**, **E-STR3**). Some estimators were able to accommodate predictors of missing outcome (i.e. **E-IV6**, **E-IV7**, **E-STR3**, **E-IV8**).

When treatment receipt was binary, the main analysis used an estimator that used all available data, allowed the inclusion of baseline predictors of outcome, and enabled

a relaxed MAR assumption. Other estimators explored the sensitivity of estimates to changes in these model specifications. The main difference was whether or not estimators made use of outcome data from participants whose adherence measures were missing. The effectiveness analyses showed that effect estimates were greater for the treatment fidelity assessment sample than the full trial sample. This was particularly true at nine months (when the estimate was significant) and 15 months after randomisation. These effects were likely an artefact of the MCAR assumption about HbA_{1c} data that were discarded. This also provides an explanation as to why the 2SLS estimates were larger than the other efficacy estimates. Estimators that allowed the inclusion of baseline predictors of outcome tended to demonstrate greater precision than those that did not. This would seem to be particular important in the estimation of efficacy, given the unavoidable loss in precision in comparison to estimators of effectiveness. Finally, comparing precision between the IV and stratification models, the latter's standard errors tended to be slightly smaller than those of the IV models. This was most likely a consequence of the fact that these models included predictors of membership of the principal strata. All of these models made the assumption that treatment was appropriately categorised. This is to assume that all participants whose treatment was below a certain level did not receive D6 intervention and all those whose treatment was above that level did receive it. If this categorisation were false, this would undermine the exclusion restriction.

When the treatment receipt variable was continuous (using MITI Global Spirit), the predicted treatment effects at the maximum difference in potential dose between the counterfactual situations decreased with time. The estimated treatment effect at nine months after randomisation was statistically significant, but not in the direction of treatment benefit. This was consistent with the pattern of effectiveness estimates for the treatment fidelity assessment sample. It suggested that the discarding of outcome data for those with missing adherence measures led to bias. The model with continuous treatment receipt assumed that MITI Global Spirit provided an accurate measure of the exposure variable and that the relationship between the difference in potential dose and the treatment effect was linear. This second assumption can only be tested with a much larger sample. The utility of predicted treatment effect at the maximum difference in potential dose may be limited as this applies to the small or perhaps non-existent subpopulation who would receive no psychological treatment if offered control and a full version of the treatment under its offer.

All of the models fitted in this chapter provided randomisation-based estimators. This means that comparisons were between groups that were defined by the outcome of random treatment allocation. All of these estimators assumed that there was no treatment effect amongst the latent classes of never takers and always takers. Put another way, this assumed that there was no effect of treatment offer on outcome other than through receipt of active intervention when offered it. This assumption is not empirically testable but may be vulnerable in a trial where clinicians and patients could not be kept blind, such as the D6 study. A limitation of this research was the methodology that I used to measure treatment receipt. As described in Chapter 3, I sampled from and rated one or two audio recordings of treatment per patient from sessions two, three and four of a treatment that could comprise as many as 12 sessions. The sample did not cover all trial participants and it is also possible that this sample did not provide a full picture of treatment fidelity over the whole course of the treatment regimen. In addition to this, the ratings of treatment fidelity did not cover other important aspects of process evaluation such as number of sessions attended and therapist alliance.

The lack of a treatment effect in the D6 trial adds to the mixed evidence for diabetes-specific, nurse-delivered psychological therapy that exists in the literature. This research is consistent with the findings of Jansink et al. (2013a), who tested nurse-led structured diabetes care (which included nurses being trained in MI) in comparison to usual care in the general practice setting. They found no evidence of an effect of this intervention on levels of HbA_{1c}. This is in contrast to Ismail et al. (2008) who found that motivational enhancement therapy plus CBT in comparison to usual care led to a reduction in HbA_{1c} of 5.0 mmol/mol for patients with type 1 diabetes.

The effectiveness analysis concluded that there was no effect of the *offer* of treatment while this analysis found that there was no effect of its *receipt*. The fidelity analysis described in Chapter 3 found that primary care nurses may not be suited to the acquisition and delivery of psychological skills, despite intensive training. This chapter has concluded that there was no evidence of a benefit of receipt of nurse-led psychological treatment on HbA_{1c} amongst people with poorly controlled T2D.

Chapter 8

General discussion

8.1 Overview of thesis

This thesis has explored many aspects of the problem of treatment contamination in trials of complex interventions in mental health. It started by investigating the processes driving contamination in this area of medical research, its quantity, the steps researchers take to minimise the problem, and the extent to which its avoidance impacts on outcomes. In its early stages it also assessed the extent of the problem in a particular trial, D6, that motivated the research. The trial assessed the effectiveness of psychological treatment for people with poorly controlled T2D. The trial was chosen to motivate this research because anecdotal evidence and the primary treatment fidelity assessment suggested that there may have been some treatment receipt amongst patients in the control arm. The research provided an opportunity to assess treatment fidelity for a large subsample of participants and then to estimate the effect of treatment receipt on outcome.

This research has focused throughout on the estimation of efficacy in the context of a trial with contamination, although these methods can easily be adapted to address treatment non-receipt in the active intervention arm (non-compliance) as well. I summarised existing randomisation-based efficacy estimators for evaluating this in individual randomised trials. I then developed a novel estimator of efficacy in trials with contamination and non-compliance measured on a continuous scale. These estimators were later applied to D6 for a secondary (efficacy) analysis of the trial's primary outcome measures. I used Monte Carlo simulation to compare two design options for addressing the problem of contamination (one of which utilised the aforementioned efficacy estimators). An online decision support tool that utilises the results was developed and published.

8.2 Main findings

8.2.1 Primary research objective

The primary research objective was to compare the efficiency of two competing trial design options that address the problem of contamination and estimate efficacy. I compared cluster randomisation at the level at which contamination was expected combined with an estimator of ATE (ITT analysis) with individual randomisation along with an estimator of either CACE or ACE_{d_1,d_0} (IV analysis). The ITT estimator accounted for the clustering of outcome data that was a consequence of randomisation being applied at the level of clinician (nested clustering, as each cluster level occurs in only one treatment arm). Cluster randomisation was assumed to prevent any contamination. In the second design option a record of treatment receipt was made for all participants, enabling the estimation of treatment effect amongst those participants who would receive a greater level of treatment under its offer than they would under offer of control (treatment receipt could be on a binary or continuous scale). This estimator also accounted for clustering of outcome data caused by the fact that each clinician treated multiple participants (crossed clustering, as cluster levels occur in both treatment arms). The ITT and IV estimators provided unbiased estimates of respective efficacy estimands.

When treatment receipt was on a binary scale, the results demonstrated that there were three factors that determined the relative efficiency of the estimators under the two trial design scenarios. These were the ICC and cluster size under the cluster randomisation design, and size of the latent complier stratum (one minus the proportion of contaminators when there was no non-compliance) under the alternative design. For large proportions of latent compliers (low amounts of contamination) estimation of CACE was often more efficient, apart from at the lowest levels of simulated ICCs and cluster sizes. As the strength of clustering increased (i.e. greater levels of ICC or cluster size, or both), the estimation of CACE was progressively favoured. The findings were broadly consistent with those in the report on contamination in trials of educational interventions (Keogh-Brown et al., 2007). I described the methods of their research in detail at the end of Section 1.5.3. Briefly, that research compared required sample size in order to achieve 80% statistical power to detect a particular effect between CRCTs analysed with an ITT estimator and individual randomised trials where contamination occurred with an estimator of CACE. They did not describe the contamination process that they mimicked.

However, its effect can be presumed to be similar to the one I have investigated, i.e. participants either receiving active or control treatment when allocated to control. They found that when cluster size was 10 and ICC was 0.04, a greater sample size was required for the CRCT design option up to a level of contamination in the individual randomised trial design option that was just under 30%. In other words, at these magnitudes of cluster size and ICC, and when contamination was less than this level, the individual randomised trial design (with estimation of CACE) was more efficient. Likewise, I found that when cluster size was 10 and ICC was 0.05 (these levels being closest to those used in Keogh-Brown et al., 2007)), the individual randomised trial design was more efficient up a level of contamination that was between 20-30%. Keogh-Brown et al. (2007) found that when cluster size was 30 and ICC was 0.04, the required sample size was greater for the CRCT option at all levels of contamination that they investigated (i.e. the amount of contamination at which the individual randomised trial required more participants was some level greater than 30%). In comparison, I found that when cluster size was 20 and ICC was 0.1, the individual randomised design option was more efficient up to between 50-60% contamination.

When there were both treatment contamination and non-compliance (i.e. more non-adherence), there were fewer combinations of ICC and cluster size under which estimation of CACE was more efficient. This was because the IV approach was weaker than before – the presence of never takers reduced the proportion of latent compliers, thereby making the estimation of CACE less precise at a given level of contamination. I found that neither sample size nor treatment effect size were directly related to the relative efficiency of the two design options.

When treatment receipt was on a continuous scale, relative efficiency between the two design options was driven by strength of clustering under the cluster randomisation design, and the size of the dose complier stratum and magnitude of dose compliance within this subpopulation under the alternative design. The magnitude of dose compliance parameter represented the size of the counterfactual difference in treatment receipt between the offers of treatment and control amongst those participants where this difference would be positive (I named this subpopulation dose compliers). When all control participants received a dose of zero and all active intervention participants received a full dose, the relative efficiency results were similar to those found when treatment receipt was binary. This was because the underlying conditions between binary

treatment receipt and continuous treatment receipt with all participants either receiving full dose or no dose are equivalent. As the magnitude of dose compliance became weaker, the efficiency of the cluster randomisation design option (clusters defined by level at which contamination occurred and estimation of ATE) was favoured under more levels of the other relevant parameters. The reason for this was that as dose compliance became weaker, the dose compliers were receiving greater doses of treatment under offer of control and lower doses of treatment under its offer. In effect, this meant more treatment contamination and non-compliance, which favoured the cluster randomised design option.

8.2.2 Secondary research objective

The secondary research objective was in two parts. The first was to summarise and develop randomisation-based estimators of efficacy in an individual randomised trial with contamination. The second was then to apply these methods to the analysis of the D6 trial, having previously constructed individual-level measures of treatment receipt. When treatment receipt is measured on a binary scale, an efficacy estimand that has received a large amount of attention in the literature is CACE. This is the effect of treatment offer amongst those participants who would receive treatment when offered it and would not receive it when offered control. When treatment receipt is measured on a continuous scale, I defined the estimand $\text{ACE}_{\text{d}_1,\text{d}_0}$ as the effect of treatment offer amongst those participants who would receive some dose of active treatment under offer of treatment (d_1) and another dose of active treatment under offer of control (d_0).

I summarised two main analysis approaches for estimating CACE: stratification estimators (using the principal stratification framework) and IV estimators. The estimators are unbiased under similar sets of assumptions (more on this in Section 8.3.3). One relative strength of the stratification estimator over use of IVs is that it enables the assumption regarding the missing data generating process to be further relaxed, in particular allowing latent compliance status to predict missingness. One drawback of this estimator is a lack of commands or functions in the most commonly used statistical software packages that allow its calculation (the estimation must be performed using a mixture model in specialist structural equation software such as MPlus).

I described the development of a novel estimator of treatment effect in a randomised trial with a continuous measure of treatment receipt in both the treatment and control

arms. This work built upon the analytical methods proposed by Maracy and Dunn (2011). The new estimator makes four assumptions. Firstly, exchangeability of potential dose and potential outcome between levels of random treatment allocation. Secondly, independence of treatment allocation and the difference in the error terms between counterfactual outcomes. When there is no difference between potential dose, this is the exclusion restriction and, when there is a difference, this assumes that there is no unaccounted variability in ATEs within levels of latent compliers. Thirdly, monotonicity, which is to assume that nobody would receive a greater dose under offer of control than treatment. Finally, a linear dose-response relationship. Under these assumptions, the causal estimand can then be estimated using IV methods. The interpretation of this parameter is the causal effect of treatment on outcome that is associated with a one-unit increase in the difference between the counterfactual doses. The parameter may be easier to interpret when it is converted into an effect that is determined by a particular difference between the counterfactual doses, for example the effect associated with full dose.

I applied these estimators of efficacy, for continuous and binary measures of treatment receipt, to the D6 trial. The results showed no evidence of an effect of treatment receipt on outcome. I demonstrated a number of approaches that can be taken towards missing data assumptions. However, the results showed little sensitivity to these assumptions, which was likely related to the finding that there was no relationship between treatment receipt and missingness.

8.2.3 Tertiary research objective

The tertiary research objective was to review problems and solutions associated with contamination in trials of complex interventions in mental health research. I conducted a large and comprehensive scoping review of contamination in this setting, which was the first of its kind in this area of medicine. The results of the review showed that those designing trials perceived five main processes leading to contamination. In a large majority of the trials, contamination was driven by two processes in particular. The first was staff delivering the active intervention to patients in the control arm (as simulated in Chapter 5). This might be due either to a given clinician delivering both active and control treatments or to a clinician who is not involved in providing active treatment learning details of the intervention and passing this on to control participants. The second major process was communication between trial arms, which could be between clinicians

or participants. Out of a total of 238 trials, I found 25 that measured and reported treatment receipt on a binary scale in the control arm. The mid-point in the reported levels of contamination was 13% (IQR 5-33%; range 0-72%). The results suggested that the problem in mental health trials may be more limited than some researchers and funders expect. Having said this, the range suggests that there may be some trial designs, interventions, or populations where it is of particular concern. The review found that those designing trials use a variety of trial conduct solutions to minimise or prevent contamination. Common examples were to design the trial so that each clinician was responsible for providing a single treatment, monitoring the treatment being delivered to control participants, and holding treatment sessions at different times or in different places, amongst others. The review featured four trials that had each allocated treatment at both cluster and individual levels. I obtained treatment effect estimates for subtrials that were defined by the level at which treatment was allocated. For each trial I then compared the magnitudes of effect sizes between the subtrials. I found no evidence of a difference, which suggested the following possibilities: cluster randomisation did not prevent contamination, the anticipated amount of it had been overstated, or cluster randomisation led to a similar degree of bias to that caused by contamination.

8.3 Implications for trials

8.3.1 Statistical design

The results from this research provide those designing trials with a tool for choosing the more efficient statistical design option to address the problem of treatment contamination when estimating efficacy. In order to provide researchers with this information, the results from the Monte Carlo simulations in Chapter 5 have been placed online at `https://nicholasmagill.shinyapps.io/shiny_app/`. The findings build on those of Keogh-Brown et al. (2007), who also compared the same design options that were assessed in this research. These results can be summarised further and combined with those of a comparison of two methods for addressing contamination when estimating effectiveness (from Slymen and Hovell, 1997), in order to provide a general set of guidelines for tackling the problem through trial design. The two design options that were compared by Slymen and Hovell (1997) were:

A. Random treatment allocation at the cluster level combined with estimator of ACE (ITT analysis) that accounts for clustered data,

B. Random treatment allocation at the level of the individual, inflation of sample size to account for the statistical power lost due to the estimated effect size being biased towards the null, and use of estimator of ACE (ITT analysis).

Design option B was described in Section 1.5.1. Slymen and Hovell (1997) compared sample size requirements under these two designs at a number of levels of cluster size, ICC, and contamination. In other words, they assessed the relative efficiencies of the methods as a consequence of the costs (in terms of statistical power) of cluster size and ICC (design option A) and contamination (design option B). They calculated sample size requirements for a trial with a continuous outcome for the two design options as follows:

A. $N' = N[1 + I(k-1)]$, where N is required sample size when randomising individuals,

B. $N^* = N/p$, where N is required sample size when randomising individuals and p is the proportion of control participants receiving the active treatment.

In their results they calculated the ratio between the required sample sizes (N'/N^*) at levels of cluster size (2; 10; 30; 100; 500), ICC (0.05; 0.1; 0.2; 0.4) and contamination (10%; 20%; 30%; 40%; 50%). They then summarised this by cluster randomisation inflation factor ($D := 1+I(k-1)$; described in Section 1.5.1) and amount of contamination (Table 1 of Slymen and Hovell, 1997). I have used those results to calculate, for a particular level of contamination, the approximate level of the inflation factor (to one decimal place) above which it was more efficient to randomise individuals and inflate sample size. This is summarised in the left half of Figure 8.1. In the same figure I have summarised the results from this research in terms of the level of the inflation factor at which it became more efficient to randomise individuals and estimate CACE as opposed to using cluster randomisation (and estimating ATE) for given levels of contamination. This is shown in the right half of the figure.

The figure shows that, for both the estimation of effectiveness and efficacy, when contamination is low (i.e. 10%) the cluster randomisation option is favoured under modest inflation factors, i.e. low levels of clustering. As the amount of contamination increases, so does the statistical cost of the individual randomisation options, and therefore the range of inflation factors at which cluster randomisation is favoured increases. The range of inflation factors within which the cluster randomisation option is favoured is smaller for the estimation of efficacy than for effectiveness at given levels of contamination. This

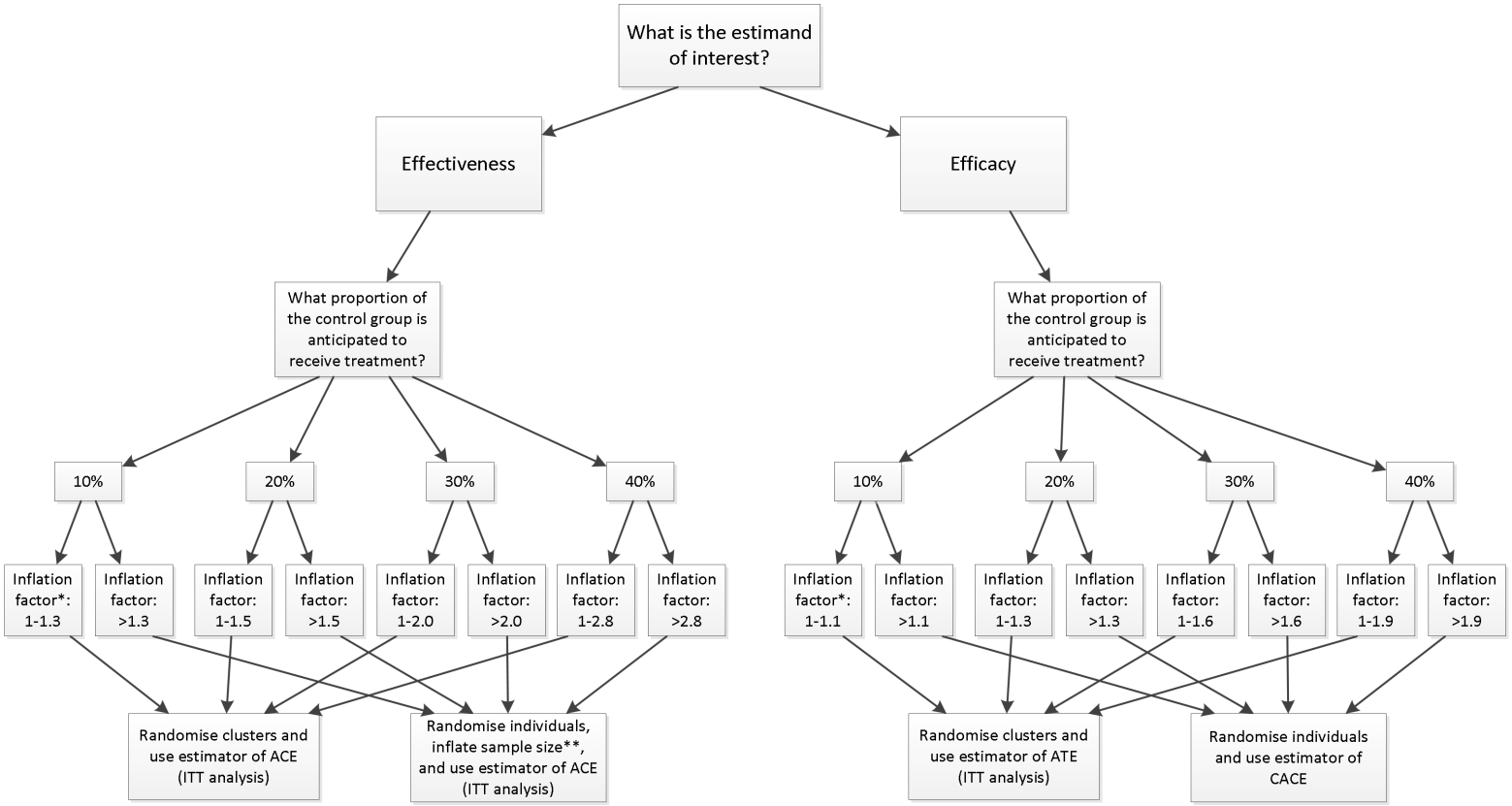


Figure 8.1: Flowchart demonstrating the more efficient trial design method for the estimation of either effectiveness or efficacy in trials with contamination (no non-compliance). Recommendations of trial design options for estimation of effectiveness are from Slymen and Hovell (1997) and for evaluation of efficacy are from this research project (results described in Chapter 5 and implemented as the decision tool in Chapter 6). The cluster randomisation design options (for estimation of effectiveness and efficacy) are assumed to prevent contamination entirely. Contamination is assumed to be measured on a binary scale. *Inflation factor $= 1 + I(k-1)$; **Inflation of sample size to recover power lost due to contamination $= 1/p$, where p is proportion of participants in the control arm who are expected to receive the treatment.

implies an extra efficiency cost of inflating the sample size in an individual randomised trial and estimating ACE compared to a trial with individual randomisation and estimation of CACE.

Throughout this thesis I have focused on trials in the field of mental health research. However, these findings regarding optimal statistical design of trials with expected contamination have utility in other areas of medical research. This is demonstrated by the fact that the research of Slymen and Hovell (1997) was in the context of adolescent consumption of tobacco and alcohol. It is also illustrated by the areas of medical research where interventions have been shown to be at risk of contamination. In Chapter 1 I described relevant literature in the domains of educational intervenions, geriatric medicine and cancer screening. As a consequence, the findings I have described regarding the design of trials targetting efficacy may have some application in these areas.

8.3.2 Conduct

Any sample size inflation using the statistical design options above will lead to an increase in the financial cost and duration of the research. Another set of design options that those researchers designing trials should consider are trial conduct methods. These are design options that affect the implementation of the trial (e.g. type of treatment, arrangements for its delivery, recruitment criteria) and can be used specifically to prevent delivery of treatment within the control arm. They comprise design methods that are straightforward in comparison to those of statistical design and, if effective, may reduce or negate the need to alter statistical design. This research aimed to describe the conduct methods that trials in mental health take in order to address the problem. Very little literature has assessed these options previously, with the survey of Keogh-Brown et al. (2007) the only notable example.

The main recommendation is that researchers must first consider the process (or processes) driving contamination before planning what to do about the problem. There is some evidence that trialists are already doing this (in mental health trials at least) given that the trial conduct solutions that were most commonly used related to the two contamination processes that were found to occur most frequently: staff delivering the active intervention to control participants and communication between trial arms (see Table 2.4 in Chapter 2).

The summary and description in this research of trial conduct solutions for minimising

contamination is the first time this has been attempted. Researchers are advised to consider these in the design stage of a trial. The caveat is that the findings do not provide any information as to the effectiveness of these conduct solutions in preventing contamination.

8.3.3 Analysis

A requirement for any efficacy analysis of a trial with contamination is that there must be a measure of treatment receipt for participants in the control arm. In practice, many trials where contamination is expected to be a problem will also encounter treatment non-receipt in the intervention arm, implying that the previous statement is true for the whole trial. The ease of doing this depends on the study. For example, in D6 the evaluation of treatment fidelity involved assessing the content of treatment sessions. This was labour intensive and required specialist skills. However, assessments of treatment receipt may be much simpler than this, for instance measurement of therapy session attendance or asking participants whether they received a particular intervention outside the realm of the trial.

It is recommended that the estimation of efficacy should avoid the use of as-treated or per protocol analyses, both of which must make very strong assumptions about absence of selection bias in order for the estimator to be unbiased. Instead it is recommended that researchers consider estimators of CACE. This estimand can be evaluated under a set of assumptions that are more plausible than the as-treated or per protocol analyses. Of these assumptions, two may be problematic in the context of contamination in mental health trials. These are the assumption of no interference and the exclusion restriction. The breaking of the first of these is sometimes referred to as spillover (VanderWeele, 2015). This may happen when an outcome is affected by communication between individuals, which I found to be one of the more prominent processes leading to contamination. There is a large and developing literature on causal estimands and estimators in this field (Tchetgen and VanderWeele, 2012; Vanderweele et al., 2013). The notation becomes more complicated than described in this research because potential outcomes now depend on the exposure status of more than one individual. There are multiple estimands that can be considered, for example the direct effect of treatment on outcome whilst holding others' exposure status constant. Another example is the spillover (indirect) effect of another's treatment on a particular individual's outcome whilst holding this individual's exposure constant. There is a difference between these estimands and the parameter that

I have been considering. The estimands from the spillover literature are interested in the effect of the offer of treatment on outcome. When the assumption of no interference is broken, this parameter is then partitioned into direct and indirect effects. The efficacy parameter that I have been considering is simply the effect of treatment offer amongst those individuals who would receive it if offered and would not receive it if offered control. Specifically, it is concerned with the total effect of treatment offer within a subpopulation defined by participants' treatment receipt, i.e. the sum of direct and indirect effects of treatment within these individuals.

The other assumption which may be problematic is the exclusion restriction. This may be vulnerable in unblinded trials, which are common in mental health research. It is plausible to imagine the offer of treatment alone affecting some participants' behaviour in a manner that may also be related to outcome. A documented example of this is "resentful demoralisation" where trial participants could become frustrated by not being offered their desired treatment (Onghena, 2005). Those in the control group are often cited in this context. It is realistic to imagine that this feeling could lead to a change in behaviour (e.g. healthier lifestyle) that could affect outcome. On a separate point, incorrect categorisation of continuous treatment receipt could also break the exclusion restriction.

More generally, I note that I have found no examples of mental health trials that have addressed contamination using an analysis of CACE, unlike other areas of medical research.

8.3.4 Reporting

A recommendation from this research is that trials should report more fully the processes leading to contamination, the quantity of it, and a description of what was done to ameliorate the problem than is commonly done at present. In the scoping review of problems and solutions associated with contamination only 12% of trials that were reviewed reported treatment receipt in the control arm, and many were vague about the process that was considered to be driving the problem. Further methodological research on the problem will remain a challenge unless these aspects of trials are better reported. This is likely to be true not only in mental health but also in other research areas where contamination is a problem, for example educational interventions and disease screening.

Solutions to the sparseness of reporting partly stem from expectations regarding the

content of publications (e.g. the CONSORT guidelines) but also relate to the ways in which trials are designed to assess treatment receipt. On the first of these, there are currently two relevant CONSORT extensions that cover contamination, one for cluster randomised trials (Campbell et al., 2012) and one for trials of nonpharmacological treatments (Boutron et al., 2017). At present the CONSORT extension for cluster randomised trials includes one item that requires authors to state the rationale for the use of cluster randomisation. This is followed by two examples of the use of cluster randomisation to prevent contamination. The item could perhaps be expanded to include a requirement for an explanation of how contamination is expected to occur, if this is part of the rationale for randomising clusters. The CONSORT extension for nonpharmacological treatments includes an item on "details of whether and how adherence of participants to interventions was assessed or enhanced" (item 5d; Boutron et al., 2017, p. 43). This item could be updated to include specific mention of whether and how adherence of participants to the control treatment was assessed.

The second recommendation regarding the reporting of contamination relates to how trials are designed to assess treatment adherence. At present many trials evaluate treatment receipt amongst a small number of participants in only the active intervention arm and make the assumption that treatment receipt in the control arm is not possible. It is recommended that, where efficacy is the causal estimand and resources allow, adherence should be assessed in both treatment and control arms, preferably for a large sample of participants.

8.4 Contribution to statistical methods for dealing with contamination and non-compliance

I have summarised existing statistical methods for addressing the problem of treatment non-adherence in RCTs. This has included the defining of target treatment effects and approaches to estimating these when treatment receipt is on a binary scale. I have specified the conditions under which estimators of these effects are causal. I have also described methods that account for clustering of outcome data and approaches for addressing assumptions surrounding missing outcome values. My novel contribution was the development of an estimand for the causal effect of treatment offer amongst a subpopulation who would receive some dose of active treatment under its offer and some dose of active treatment under offer of control. I explained the necessary assumptions for

its calculation and described how it could be estimated. I then demonstrated best practice for estimation of causal estimands in the presence of binary or continuous measures of treatment receipt by applying these to the D6 trial.

The statistical methods I have explored for dealing with contamination and non-compliance have focused on a particular set of circumstances. The methods apply to superiority trials with two arms where participants receive either the experimental or comparator treatment (or perhaps some dose of the active treatment). They have been restricted to studies with a continuous outcome. A large part of the causal inference literature has placed the same restrictions on methods for estimating efficacy in trials with non-adherence. However, some progress has been made in relaxing these. For example, Fischer et al. (2010) developed structural mean models for estimating efficacy in trials with two active treatments have Gillespie et al. (2015) have applied these to non-inferiority trials. They demonstrated how the 2SLS method can be used to estimate efficacy. In particular, they fitted a model to estimate treatment effects within the intervention and comparator arms, where the difference between these effects was the estimate of efficacy. The final step was to assess whether the confidence interval for the efficacy estimate was within the non-inferiority margin. Other progress has been made in applying randomisation-based efficacy estimators to binary outcome data (Cook et al., 2018). The method used the two-stage residual inclusion approach and was demonstrated in a surgical trial with non-adherence. The first stage of this method is to regress treatment received on random allocation and save the residuals. The second stage is to use a Poisson model (for estimation of relative risk) and regress binary outcome on treatment allocation and a covariate representing the residuals from the first stage. This provides an estimator of the risk ratio of treatment receipt on outcome that adjusts for any confounding of this target pathway. Elsewhere, Clarke and Windmeijer (2010) explored how structural mean models could be used to identify LATEs in trials with binary outcomes. They summarised the required assumptions for estimating such effects with additive and multiplicative models.

Methods have been developed for estimating efficacy in trials with treatment switching (participants moving between trial arms) in the context of time-to-event data. The challenges here are to model the outcome as well as the treatment switching, which is treated as censoring (i.e. any participant's follow-up after the switch is excluded; Watkins et al., 2013). Inverse probability weightings can be used to adjust for censoring, i.e. those participants who are not censored are assumed to be representative of those

who are censored (conditional on model covariates) and are therefore upweighted. This assumption rests on there being no unmeasured confounders of the relationship between switching and potential time to event when there is no switch. The weights are then used in the model for the outcome. In addition, other methods have been proposed for applying IV methods to censored outcome data. For example, Tchetgen et al. (2015) developed methods that are analogous to the 2SLS and ATR (control function) methods that I have described. In the context of survival outcomes, these methods involve using predicted values or residuals in the additive hazard outcome model.

Many of the extensions to methods for estimating efficacy in trials have been developed in recent years. My research provides a contribution to this set of methods with particular application to the analysis of trials in mental health.

8.5 Contribution to the field of diabetes treatment in the context of psychological medicine

Type 2 diabetes is a chronic disease with serious implications for patients and the organisation of healthcare services. For patients, complications can include heart disease, stroke and diabetic retinopathy; the disease is associated with a reduction in life expectancy of 10 years (Melmed et al., 2016). For a healthcare provider such as the NHS the costs of treatment of chronic diseases are potentially large, especially T2D given that its prevalence in the UK increased threefold to 4.5% between 1991-2013 (Hillier and Pedula, 2001). The better management of chronic diseases, in particular outside hospital, is considered a priority for the NHS (Goodwin et al., 2010). In this context the D6 trial tested the effect of nurse-delivered psychological treatment within the setting of primary care for people with T2D and poorly controlled blood glucose. This research project has made contributions to this field through the assessment of treatment fidelity amongst a large sample of trial participants and an efficacy analysis of the effect of treatment receipt on primary outcome (HbA_{1c}).

The treatment fidelity assessment (Chapter 3) showed that after MI and CBT skills training, nurses had basic competencies in some psychological techniques, although there also seemed to be some delivery of psychological treatment by nurses in the control arm. This built upon the findings of a smaller fidelity assessment that was carried out as part of the primary analysis of D6 (Ismail et al., 2018). That assessment sampled therapy session

recordings by nurses rather than participants and rated only a 20-minute window within each one. Group differences were smaller in comparison to the assessment reported here and often did not reach statistical significance. The findings of the assessment with the larger sample demonstrated the merits of sampling recordings by participants and rating the whole of each session. Similar RCTs should assess treatment fidelity in a large sample of participants and should evaluate both treatment receipt in the intervention arm and the absence of intervention in the control arm. This enables an assessment of what treatments participants received and therefore allows further analysis of treatment receipt and mechanism.

There were many factors that may have contributed to limited development in skills, including individual nurse characteristics and organisational factors such as lack of support and appropriate surgery infrastructure (Graves et al., 2016). Future studies should focus on selection strategies for nurses that maximise chances of success, enhance the training of nurses, give further consideration to the choice of the comparator treatments of standard care and attention control, or contemplate the possibility that primary care nurse acquisition of high-level MI and CBT skills is not a viable approach to improved self-management among diabetic patients with persistent suboptimal control.

The main conclusion of the efficacy analysis (Chapter 7) was that no evidence of an effect of treatment receipt on glycated haemoglobin was observed. This was based on a definition of treatment receipt that stemmed from the assessment of primary care nurses' delivery of psychological skills. It was also observed that participants in both trial arms demonstrated a modest improvement in glycaemic control. This may have been related to participants in both trial arms receiving more contact time with nurses. A possible explanation for the concurrence of lack of efficacy and the small improvement across both trial arms was the combination of characteristics of the trial population as a whole. The population, whose median duration of diabetes was nine years, may not have been capable of response to intervention and further improvement in glycaemic control. Such a population had most likely already been offered considerable medical input and might have included people who do not engage well with services.

In summary, primary care nurses struggled to acquire and deliver psychological skills such as MI and CBT to a high level, despite the use of an intensive, manualised training programme with ongoing supervision by an experienced clinical psychologist. The effectiveness and efficacy analyses showed no effect of treatment on glycated haemoglo-

bin. Further studies may be needed to determine whether, for patients to benefit from such therapies, a different skill set may be needed in the healthcare professional or a re-organisation of nurse practitioner time to allow for greater engagement in training and delivery.

8.6 Strengths and limitations

The research described in this thesis constitutes a comprehensive study of the challenges of and methods for addressing the problem of contamination in RCTs in mental health and includes novel methods and findings. On the primary research objective, it is the most extensive comparison of two trial design options for addressing contamination to date. It is the first to provide a decision support tool for researchers who are considering what do about the problem. On the secondary research objective, the work has described a novel estimator of efficacy in the context of a trial with continuous measures of treatment receipt in both treatment and control arms. It is the first study that I am aware of that has used randomisation-based estimators of efficacy in the presence of contamination in a mental health trial. On the tertiary research objective, it comprises the largest review of contamination in trials and the first time this has been done in mental health.

The research has some notable limitations, generally with respect to the types of trials that are being considered and the definition of contamination. There were also more specific weaknesses in terms of the Monte Carlo simulations and the trial that was chosen as a motivating example throughout. The research has defined contamination as the receipt of intervention amongst participants in the control arm. This definition could be broadened to include contamination between experimental intervention arms, for example in a trial testing two active treatments. In addition, this research has only considered superiority trials. This could be extended to non-inferiority trials, where the effects of contamination would be to provide false evidence of equivalence between treatments.

The data simulations were limited to one type of contamination, namely the crossover of staff from treatment to control arms. I did not simulate trial data under the other main process that can drive contamination, which is when it occurs due to communication between patients or clinicians between different trial arms. The main challenge of doing this would be to mimick realistically the spread of information between and

within clusters. The investigation of bias in cluster compared to individual randomised trials in the report of contamination in educational interventions attempted to do this (Keogh-Brown et al., 2007). However, the authors took a "common sense" rather than empirically-based approach to selecting levels for relevant parameters. It is not clear how realistic levels could be selected for parameters representing the diffusion of information. On a separate point, the comparison with the Keogh-Brown et al. (2007) research highlights a strong assumption that I made in the data simulations: the lack of bias associated with cluster randomisation. This could be investigated in future research. Finally, on the subject of the data simulations, a weakness of the research is arguably the comparison of ATE and CACE. These estimands are distinct as demonstrated by the populations they apply to. An estimator of ATE applies to the whole population, whereas one of CACE can be said to apply to the complier subpopulation. The justification for comparing these target parameters was that they both constituted valid approaches to efficacy under the respective designs of cluster and individual randomisation. However, some investigators might take the view that ATE is the more general estimand and hence prefer cluster randomisation and the ITT estimator irrespective of efficiency arguments.

There were some limitations regarding the design of the data simulations. First, I chose a data generating mechanism which included a cluster-level random effect for outcome at baseline and a random effect at the level of the cluster (therapist) for outcome at follow-up. These effects reflected clustering due to participants sharing environments and therapists, respectively. The baseline variable was included in analyses, thereby adjusting for baseline cluster effects. It would have been possible to simulate clustering in other ways by, for example, including a cluster-specific random effect that influences outcome at both baseline and follow-up for an individual from a particular cluster. In this case, it would have been necessary to adjust each individual's follow-up outcome for the baseline cluster mean as well as the individual's baseline outcome. I could also have considered more than two levels of clustering, for example the effect of shared environment in addition to therapist clusters (where therapists are nested within communities) on outcome after treatment. I specified a moderately simple data generating mechanism in order for the methods and findings to be realistic and useful to those planning trials. There may be some interest in investigating other data generating mechanisms in the future according to specific trial designs and scenarios in which contamination is a problem. Second, I simulated some trial scenarios that were not of interest to trialists under the trial design process for addressing contamination that I recommended. As a

reminder, my suggested plan is first to calculate sample size for an individual randomised trial with no contamination according to particular levels of treatment effect size, power and significance. Following this, the simulation results reported here should be used to compare statistical efficiency between the design options associated with the presence of cluster randomisation or contamination. In the data simulations I generated data under many trial scenarios, which were defined by the levels of various input parameters. These scenarios were balanced due to the fact that data were generated at combinations of every level of these parameters. This meant that there were some trial scenarios that were implausible according to the planning process recommended above. For example, a researcher would never plan a trial with power of 80%, a small standardised effect size (of 0.2) and a sample size of 100. Although some of these trial scenarios are not practicable, the generation of data under all scenarios allowed me to plot complete isosurfaces for the equivalence of efficiency between design options.

D6 was chosen as the motivating example because of anecdotal evidence of treatment contamination. This happened despite the trial being cluster randomised, a design that was chosen in part to prevent treatment receipt in the control arm. The trial was also chosen because it enabled a measure of treatment receipt to be collected for a large subsample of participants. In reality, such a measure was generated for only 151 out of 334 participants (45.2%). Of these, there existed data for only 107 participants on both treatment receipt and outcome at 18 months after randomisation (the primary endpoint). The fact that the sample was small led to estimates of efficacy with large standard errors. There were a number of aspects to the efficacy analysis of the trial that suggested some uncertainty that the estimators calculated the causal effect of treatment on outcome. In particular, assumptions remain that there was no effect of treatment offer amongst never and always takers (the exclusion restriction) and, for the continuous measure of treatment receipt, that the correct functional relationship between dose and treatment effect was linear. The first of these, which is untestable, may be vulnerable in an unblinded trial such as D6. This is because of the possibility that the offer alone of psychological treatment may influence outcome by some route other than receipt of treatment, for example by prompting participants to make better use of usual care or improve aspects of lifestyle such as diet and exercise. In addition, there is uncertainty regarding the correct measure of treatment receipt. If the true form is continuous then its dichotomisation will result in the breaking of the exclusion restriction. Finally, without a much larger sample size, it is impossible to assess the true shape of the relationship

between treatment receipt and effect.

8.7 Future work

The limitations of the work that addressed the primary research objective provide some clear avenues for further research. For instance, there would be utility in extending the scope of the data simulations to include the scenario where contamination is driven by communication amongst patients or clinicians in different trial arms. This would require some information about the extent and speed of the spread of information from the treatment to the control groups and subsequently between participants within the control group. I am not aware of any information within the literature on this. It may be necessary to use surveys of trialists or even a pilot study to investigate relevant simulation parameters and their likely levels. As mentioned earlier, future data simulations investigating the relative efficiency of trial design options for addressing contamination should consider modelling the biases associated with cluster randomisation. These biases have been well described in the past (Puffer et al., 2003; Hahn et al., 2005). Finally, it would be informative to compare the financial costs of the various approaches for addressing contamination in trials that are designed to estimate either effectiveness or efficacy. This is likely to be a concern for funders but there is currently no information in the literature on this subject.

The work on efficacy estimators has highlighted some potential extensions to the methods. The estimands that I have considered relate to treatment efficacy, which was defined as the effect of treatment receipt amongst either the whole population or the compliers. One possible extension in the context of contamination would be to explore a new estimand, the contrast between offer of active treatment and receipt of control. This could be defined as the effect of treatment offer within the latent compliers and never takers:

$$\text{LATE}_{\text{c,n}} := \text{E}[Y_i(R=1) - Y_i(R=0) | T_i(1) - T_i(0) = 1, T_i(1) = T_i(0) = 0]$$

where the subscripts 'c' and 'n' refer to latent compliers and never takers respectively. The combination of latent compliers and never takers are observed as a mixture of treatment receivers and non-receivers when offered active treatment, and observed as control receivers when offered control. This target effect is a LATE and could be estimated as the average of the effects amongst compliers (CACE) and amongst never takers (assumed to

be zero), weighted by the proportions of these subpopulations:

$$\begin{aligned}\widehat{LATE_{\text{c,n}}} :=& \hat{p} \cdot \widehat{CACE} + (1-\hat{p}) \cdot 0 \\ =& \hat{p} \cdot \text{E}[Y_i(R=1) - Y_i(R=0) | T_i(1) - T_i(0) = 1]\end{aligned}$$

where $p = \frac{p_\text{c}}{p_\text{c}+p_\text{n}}$ (i.e. proportion of compliers divided by summed proportions of compliers and never takers).

On a more general level, there is a need for further research into treatment contamination in mental health trials. For instance, there is still little understanding in many cases about precisely why contamination occurs. I am not aware of a trial that has assessed and described in detail the processes that drove contamination. There may also be some utility in describing the characteristics of those participants whose treatment is contaminated. By contrast, this research has only attempted to estimate their proportion. With further methods for identifying these participants there may be some potential for reducing the problem.

References

Abroug, F., Ouanes-Besbes, L., Dachraoui, F., Ouanes, I., and Brochard, L. (2011). An updated study-level meta-analysis of randomised controlled trials on proning in ARDS and acute lung injury. *Critical Care*, 15(1):R6.

Alam, R., Sturt, J., Lall, R., and Winkley, K. (2009). An updated meta-analysis to assess the effectiveness of psychological interventions delivered by psychological specialists and generalist clinicians on glycaemic control and on psychological status. *Patient Education and Counseling*, 75(1):25–36.

Alessi, C. A., Martin, J. L., Webber, A. P., Kim, C. E., Harker, J. O., and Josephson, K. R. (2005). Randomized, controlled trial of a nonpharmacological intervention to improve abnormal sleep/wake patterns in nursing home residents. *Journal of the American Geriatrics Society*, 53(5):803–810.

Altman, D., Whitehead, J., Parmar, M., Stenning, S., Fayers, P., and Machin, D. (1995). Randomised consent designs in cancer clinical trials. *European Journal of Cancer*, 31(12):1934–1944.

Anderson, T. W. and Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, pages 46–63.

Angrist, J. D. and Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Apodaca, T. R. and Longabaugh, R. (2009). Mechanisms of change in motivational interviewing: A review and preliminary evaluation of the evidence. *Addiction*, 104(5):705–715.

Aveyard, P., Brown, K., Saunders, C., Alexander, A., Johnstone, E., Munafo, M. R., and Murphy, M. (2007). Weekly versus basic smoking cessation support in primary care: A randomised controlled trial. *Thorax*, 62(10):898–903.

Balestra, P. and Varadharajan-Krishnakumar, J. (1987). Full information estimations of a system of simultaneous equations with error component structure. *Econometric Theory*, 3(2):223–246.

Barkhof, E., Meijer, C. J., de Sonneville, L. M. J., Linszen, D. H., and de Haan, L. (2013). The effect of motivational interviewing on medication adherence and hospitalization rates in nonadherent patients with multi-episode schizophrenia. *Schizophrenia Bulletin*, 39(6):1242–1251.

Barton, M. B., Morley, D. S., Moore, S., Allen, J. D., Kleinman, K. P., Emmons, K. M., and Fletcher, S. W. (2004). Decreasing women's anxieties after abnormal mammograms: A controlled trial. *Journal of the National Cancer Institute*, 96(7):529–538.

Beck, C. K., Vogelpohl, T. S., Rasin, J. H., Uriri, J. T., O'Sullivan, P., Walls, R., Phillips, R., and Baldwin, B. (2002). Effects of behavioral interventions on disruptive behavior and affect in demented nursing home residents. *Nursing Research July/August*, 51(4):219–228.

Becoña, E. and Vázquez, F. L. (2001). Effectiveness of personalized written feedback through a mail intervention for smoking cessation: A randomized-controlled trial in Spanish smokers. *Journal of Consulting and Clinical Psychology*, 69(1):33.

Bernstein, G. A., Layne, A. E., Egan, E. A., and Tennison, D. M. (2005). School-based interventions for anxious children. *Journal of the American Academy of Child & Adolescent Psychiatry*, 44(11):1118–1127.

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review*, 8(2):225–246.

Borland, R., Balmford, J., and Benda, P. (2013). Population-level effects of automated smoking cessation help programs: A randomized controlled trial. *Addiction*, 108(3):618–628.

Borm, G. F., Melis, R. J., Teerenstra, S., and Peer, P. G. (2005). Pseudo cluster randomization: A treatment allocation method to minimize contamination and selection bias. *Statistics in Medicine*, 24(23):3535–3547.

Boutron, I., Altman, D. G., Moher, D., Schulz, K. F., and Ravaud, P. (2017). CONSORT statement for randomized trials of nonpharmacologic treatments: A 2017 update and a CONSORT extension for nonpharmacologic trial abstracts. *Annals of Internal Medicine*, 167(1):40–47.

Burgess, S., Davies, N. M., and Thompson, S. G. (2014). Instrumental variable analysis with a nonlinear exposure–outcome relationship. *Epidemiology*, 25(6):877.

Campbell, M., Fitzpatrick, R., Haines, A., Kinmonth, A. L., Sandercock, P., Spiegelhalter, D., and Tyrer, P. (2000). Framework for design and evaluation of complex interventions to improve health. *BMJ*, 321(7262):694–696.

Campbell, M. K. and Grimshaw, J. M. (1998). Cluster randomised trials: Time for improvement: The implications of adopting a cluster design are still largely being ignored. *BMJ*, 317(7167):1171–1172.

Campbell, M. K., Piaggio, G., Elbourne, D. R., and Altman, D. G. (2012). CONSORT 2010 statement: Extension to cluster randomised trials. *BMJ*, 345:e5661.

Campbell, M. K., Snowdon, C., Francis, D., Elbourne, D. R., McDonald, A. M., Knight, R. C., Entwistle, V. A., Garcia, J., Roberts, I., and Grant, A. M. (2007). Recruitment to randomised trials: Strategies for trial enrolment and participation study. The STEPS study. *Health Technology Assessment*, 11(48):1–123.

Chan, M. F., Ng, S. E., Tien, A., Man Ho, R. C., and Thayala, J. (2013). A randomised controlled study to explore the effect of life story review on depression in older Chinese in Singapore. *Health & Social Care in the Community*, 21(5):545–553.

Chanen, A. M. F., Jackson, H. J., McCutcheon, L. K., Jovev, M., Dudgeon, P., Yuen, H. P., Germano, D., Nistico, H., McDougall, E., Weinstein, C., Clarkson, V., and McGorry, P. D. (2008). Early intervention for adolescents with borderline personality disorder using cognitive analytic therapy: Randomised controlled trial. *British Journal of Psychiatry*, 193(6):477–484.

Chochinov, H. M., Kristjanson, L. J., Breitbart, W., McClement, S., Hack, T. F., Hassard, T., and Harlos, M. (2011). Effect of dignity therapy on distress and end-of-life experience

in terminally ill patients: A randomised controlled trial. *Lancet Oncology*, 12(8):753–62.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4):284.

Ciechanowski, P. S., Katon, W. J., and Russo, J. E. (2000). Depression and diabetes: Impact of depressive symptoms on adherence, function, and costs. *Archives of Internal Medicine*, 160(21):3278–3285.

Clarke, P. S. and Windmeijer, F. (2010). Identification of causal effects on binary outcomes using structural mean models. *Biostatistics*, 11(4):756–770.

Clarkson, J. E., Young, L., Ramsay, C. R., Bonner, B. C., and Bonetti, D. (2009). How to influence patient oral hygiene behavior effectively. *Journal of Dental Research*, 88(10):933–937.

Cook, J. A., MacLennan, G. S., Palmer, T., Lois, N., and Emsley, R. (2018). Instrumental variable methods for a binary outcome were used to informatively address noncompliance in a randomized trial in surgery. *Journal of Clinical Epidemiology*, 96:126–132.

Cooper, L. A., Ghods Dinoso, B. K., Ford, D. E., Roter, D. L., Primm, A. B., Larson, S. M., Gill, J. M., Noronha, G. J., Shaya, E. K., and Wang, N.-Y. (2013). Comparative effectiveness of standard versus patient-centered collaborative care interventions for depression among african americans in primary care settings: The BRIDGE study. *Health Services Research*, 48(1):150–174.

Courneya, K. S., Friedenreich, C. M., Sela, R. A., Quinney, H., Rhodes, R. E., and Handman, M. (2003). The group psychotherapy and home-based physical exercise (group-hope) trial in cancer survivors: Physical fitness and quality of life outcomes. *Psycho-Oncology*, 12(4):357–374.

Cox, D. R. (1958). *Planning of experiments*. Wiley.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., and Petticrew, M. (2008). Developing and evaluating complex interventions: The new Medical Research Council guidance. *BMJ*, 337(a1655):979–983.

Cuzick, J., Edwards, R., and Segnan, N. (1997). Adjusting for non-compliance and contamination in randomized clinical trials. *Statistics in Medicine*, 16(9):1017–1029.

Davies, M. J., Heller, S., Skinner, T., Campbell, M., Carey, M., Cradock, S., Dallosso, H., Daly, H., Doherty, Y., Eaton, S., Fox, C., Oliver, L., Rantell, K., Rayman, G., and Khunti, K. (2008). Effectiveness of the diabetes education and self management for ongoing and newly diagnosed (DESMOND) programme for people with newly diagnosed type 2 diabetes: Cluster randomised controlled trial. *BMJ*, 336(7642):491–495.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Dilley, J. W., Woods, W. J., Loeb, L., Nelson, K., Sheon, N., Mullan, J., Adler, B. L., Chen, S., and McFarland, W. (2007). Brief cognitive counseling with HIV testing to reduce sexual risk among men who have sex with men: Results from a randomized controlled trial using paraprofessional counselors. *Journal of Acquired Immune Deficiency Syndromes*, 44(5):569–577.

DiMatteo, M. R. (2004). Variations in patients' adherence to medical recommendations: A quantitative review of 50 years of research. *Medical Care*, 42(3):200–209.

Dobscha, S. K., Corson, K., Perrin, N. A., Hanson, G. C., Leibowitz, R. Q., Doak, M. N., Dickinson, K. C., Sullivan, M. D., and Gerrity, M. S. (2009). Collaborative care for chronic pain in primary care: A cluster randomized trial. *JAMA*, 301(12):1242–1252.

Dunn, G. (2013). Pragmatic trials of complex psychosocial interventions: Methodological challenges. *Epidemiology and Psychiatric Sciences*, 22(02):105–109.

Dunn, G. and Bentall, R. (2007). Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions (psychological treatments). *Statistics in Medicine*, 26(26):4719–4745.

Dunn, G., Fowler, D., Rollinson, R., Freeman, D., Kuipers, E., Smith, B., Steel, C., Onwumere, J., Jolley, S., Garety, P., et al. (2012). Effective elements of cognitive behaviour therapy for psychosis: Results of a novel type of subgroup analysis based on principal stratification. *Psychological Medicine*, 42(05):1057–1068.

Dunn, G., Maracy, M., Dowrick, C., Ayuso-Mateos, J. L., Dalgard, O. S., Page, H., Lehtinen, V., Casey, P., Wilkinson, C., Vázquez-Barquero, J. L., and Wilkinson, G. (2003). Esti-

mating psychological treatment effects from a randomised controlled trial with both non-compliance and loss to follow-up. *The British Journal of Psychiatry*, 183(4):323–331.

Dunn, G., Maracy, M., and Tomenson, B. (2005). Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: The role of instrumental variable methods. *Statistical Methods in Medical Research*, 14(4):369–395.

Ell, K., Katon, W., Xie, B., Lee, P.-J., Kapetanovic, S., Guterman, J., and Chou, C.-P. (2010). Collaborative care management of major depression among low-income, predominantly Hispanic subjects with diabetes: A randomized controlled trial. *Diabetes Care*, 33(4):706–713.

Emsley, R. and Dunn, G. (2012). Evaluation of potential mediators in randomised trials of complex interventions (psychotherapies). In Berzuini, C., Dawid, P., and Bernardinell, L., editors, *Causality: Statistical perspectives and applications*, chapter Evaluation of potential mediators in randomised trials of complex interventions (psychotherapies), pages 290–309. John Wiley & Sons.

Ersek, M., Turner, J. A., Cain, K. C., and Kemp, C. A. (2008). Results of a randomized controlled trial to examine the efficacy of a chronic pain self-management group for older adults [ISRCTN11899548]. *Pain*, 138(1):29–40.

European Medicines Agency (2018). ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials.

Fairburn, C. G. and Cooper, Z. (2011). Therapist competence, therapy quality, and therapist training. *Behaviour Research and Therapy*, 49(6):373–378.

Fischer, K., Goetghebeur, E., Vrijens, B., and White, I. R. (2010). A structural mean model to allow for noncompliance in a randomized trial comparing 2 active treatments. *Biostatistics*, 12(2):247–257.

Fischer-Lapp, K. and Goetghebeur, E. (1999). Practical properties of some structural mean analyses of the effect of compliance in randomized trials. *Controlled Clinical Trials*, 20(6):531–546.

Floyd, A. H. and Moyer, A. (2010). Effects of participant preferences in unblinded randomized controlled trials. *Journal of Empirical Research on Human Research Ethics*, 5(2):81–93.

Forchuk, C., Martin, M. L., Chan, Y. L., and Jensen, E. (2005). Therapeutic relationships: From psychiatric hospital to community. *Journal of Psychiatric & Mental Health Nursing*, 12(5):556–564.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1):21–29.

Gæde, P., Lund-Andersen, H., Parving, H.-H., and Pedersen, O. (2008). Effect of a multifactorial intervention on mortality in type 2 diabetes. *New England Journal of Medicine*, 358(6):580–591.

Gilbody, S., Bower, P., Torgerson, D., and Richards, D. (2008). Cluster randomized trials produced similar results to individually randomized trials in a meta-analysis of enhanced care for depression. *Journal of Cinical Epidemiology*, 61(2):160–168.

Gillespie, D., Hood, K., Farewell, D., Hawthorne, A., Probert, C., Stenson, R., Barrett-Lee, P., Casbard, A., and Murray, N. (2015). The use of randomisation-based efficacy estimators in non-inferiority trials. *Trials*, 16(S2):P129.

Goel, V., Cohen, M. M., Kaufert, P., and MacWilliam, L. (1998). Assessing the extent of contamination in the Canadian National Breast Screening Study. *American Journal of Preventive Medicine*, 15(3):206–211.

Goldsmith, L. P., Lewis, S., Dunn, G., and Bentall, R. (2015). Psychological treatments for early psychosis can be beneficial or harmful, depending on the therapeutic alliance: An instrumental variable analysis. *Psychological Medicine*, 45(11):2365–2373.

Goodwin, N., Curry, N., Naylor, C., Ross, S., and Duldig, W. (2010). Managing people with long-term conditions. *London: The Kings Fund*.

Graves, H., Garrett, C., Amiel, S. A., Ismail, K., and Winkley, K. (2016). Psychological skills training to support diabetes self-management: Qualitative assessment of nurses' experiences. *Primary Care Diabetes*, 10(5):376–382.

Grimshaw, J., Thomas, R., MacLennan, G., Fraser, C., Ramsay, C., Vale, L., Whitty, P., Eccles, M., Matowe, L., Shirran, L., Wensing, M., Dijkstra, R., and Donaldson, C.

(2004). Effectiveness and efficiency of guideline dissemination and implementation strategies. *Health Technology Assessment*, 8(4).

Hahn, S., Puffer, S., Torgerson, D. J., and Watson, J. (2005). Methodological bias in cluster randomised trials. *BMC Medical Research Methodology*, 5(1):10.

Hardeman, W., Lamming, L., Kellar, I., De Simoni, A., Graffy, J., Boase, S., Sutton, S., Farmer, A., and Kinmonth, A. L. (2014). Implementation of a nurse-led behaviour change intervention to support medication taking in type 2 diabetes: Beyond hypothesised active ingredients (SAMS Consultation Study). *Implementation Science*, 9(1):1.

Heirich, M. and Sieck, C. J. (2000). Worksite cardiovascular wellness programs as a route to substance abuse prevention. *Journal of Occupational and Environmental Medicine*, 42(1):47.

Hernán, M. A. and Hernández-Díaz, S. (2012). Beyond the intention-to-treat in comparative effectiveness research. *Clinical Trials*, 9(1):48–55.

Hernán, M. A. and Robins, J. M. (2018). *Causal inference*. Boca Raton: Chapman & Hall/CRC, forthcoming.

Higgins, J. P. T., Altman, D. G., and Sterne, J. A. C. (2011). Assessing risk of bias in included studies. In *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]*, chapter 8. The Cochrane Collaboration. Available from www.cochrane-handbook.org.

Hillier, T. A. and Pedula, K. L. (2001). Characteristics of an adult population with newly diagnosed type 2 diabetes. *Diabetes Care*, 24(9):1522–1527.

Hirano, K., Imbens, G. W., Rubin, D. B., and Zhou, X.-H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1(1):69–88.

Holden, S., Jenkins-Jones, S., Morgan, C. L., Peters, J., Schernthaner, G., and Currie, C. (2017). Prevalence, glucose control and relative survival of people with type 2 diabetes in the UK from 1991 to 2013. *Diabetic Medicine*, 34(6):770–780.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Howard, L., de Salis, I., Tomlin, Z., Thornicroft, G., and Donovan, J. (2009). Why is recruitment to trials difficult? An investigation into recruitment difficulties in an rct of

supported employment in patients with severe mental illness. *Contemporary Clinical Trials*, 30(1):40–46.

Imbens, G. and Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Imbens, G. W. and Rubin, D. B. (1997). Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*, 64(4):555–574.

International Conference on Harmonisation (1998). *ICH harmonised tripartite guideline for statistical principles for clinical trials*. Brookwood Medical Publications.

Ismail, K., Thomas, S. M., Maissi, E., Chalder, T., Schmidt, U., Bartlett, J., Patel, A., Dickens, C. M., Creed, F., and Treasure, J. (2008). Motivational enhancement therapy with and without cognitive behavior therapy to treat type 1 diabetes: A randomized trial. *Annals of Internal Medicine*, 149(10):708–719.

Ismail, K., Winkley, K., de Zoysa, N., Patel, A., Heslin, M., Graves, H., Thomas, S., Stringer, D., Stahl, D., and Amiel, S. A. (2018). Nurse-led psychological intervention for type 2 diabetes: A cluster randomised controlled trial (Diabetes-6 study) in primary care. *British Journal of General Practice*, 68:531–540.

Ismail, K., Winkley, K., and Rabe-Hesketh, S. (2004). Systematic review and meta-analysis of randomised controlled trials of psychological interventions to improve glycaemic control in patients with type 2 diabetes. *The Lancet*, 363(9421):1589–1597.

Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J. M., Gavaghan, D. J., and McQuay, H. J. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials*, 17(1):1–12.

Jansink, R., Braspenning, J., Keizer, E., van der Weijden, T., Elwyn, G., and Grol, R. (2013a). No identifiable Hb1Ac or lifestyle change after a comprehensive diabetes programme including motivational interviewing: A cluster randomised trial. *Scandinavian Journal of Primary Health Care*, 31(2):119–127.

Jansink, R., Braspenning, J., Laurant, M., Keizer, E., Elwyn, G., van der Weijden, T., and Grol, R. (2013b). Minimal improvement of nurses' motivational interviewing skills in routine diabetes care one year after training: A cluster randomized trial. *BMC Family Practice*, 14(1):44.

Jin, H. and Rubin, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*, 103(481):101–111.

Jo, B. and Muthén, B. O. (2001). Modeling of intervention effects with noncompliance: A latent variable approach for randomized trials. In *New developments and techniques in structural equation modeling*, chapter 3, pages 57–87. Lawrence Erlbaum Associates, Inc.

Johnson, S., Thornicroft, G., Afuwape, S., Lesse, M., Hughes, E., Waingarante, S., Miles, H., and Craig, T. (2007). Effects of training community staff in interventions for substance misue in dual diagnosis patients with psychosis (COMO study): Cluster randomised trial. *British Journal of Psychiatry*, 191:451–452.

Keogh-Brown, M. R., Bachmann, M., Shepstone, L., Hewitt, C., Howe, A., Ramsay, C. R., Song, F., Miles, J., Torgerson, D., Miles, S., et al. (2007). Contamination in trials of educational interventions. *Health Technology Assessment*, 11(43).

Khumalo-Sakutukwa, G. M., Morin, S. F., Fritz, K., Charlebois, E. D., van Rooyen, H., Chingono, A., Modiba, P., Mrumbi, K., Visrutaratna, S. D., Singh, B., Sweat, M., Celentano, D. D., Coates, T. J., and NIMH Project Accept Study Team (2008). Project accept (HPTN 043): A community-based intervention to reduce HIV incidence in populations at risk for HIV in Sub-Saharan Africa and Thailand. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 49(4):422–431.

Lamers, F., Jonkers, C. C. M., Bosma, H., Kempen, G. I. J. M., Meijer, J. A. M. J., Penninx, B. W. J. H., Knottnerus, A. J., and van Eijk, J. T. M. (2010). A minimal psychological intervention in chronically ill elderly patients with depression: A randomized trial. *Psychotherapy & Psychosomatics*, 79(4):217–226.

Lane, C., Huws-Thomas, M., Hood, K., Rollnick, S., Edwards, K., and Robling, M. (2005). Measuring adaptations of motivational interviewing: The development and validation of the behavior change counseling index (BECCI). *Patient Education and Counseling*, 56(2):166–173.

Larsen, M., Krogstad, A., Aas, E., Moum, T., and Wahl, A. (2014). A telephone-based motivational interviewing intervention has positive effects on psoriasis severity and self-management: A randomized controlled trial. *British Journal of Dermatology*, 171(6):1458–1469.

Lee, K. A. and Gay, C. L. (2011). Can modifications to the bedroom environment improve the sleep of new parents? Two randomized controlled trials. *Research in Nursing & Health*, 34(1):7–19.

Lefebvre, C., Manheimer, E., and Glanville, J. (2011). Searching for studies. In Higgins, J. P. T. and Green, S., editors, *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011)*, chapter 6. The Cochrane Collaboration. Available from www.cochrane-handbook.org.

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Lorensen, W. E. and Cline, H. E. (1987). Marching cubes: A high resolution 3d surface construction algorithm. In *ACM SIGGRAPH Computer Graphics*, volume 21, pages 163–169. ACM.

Madson, M. B. and Campbell, T. C. (2006). Measures of fidelity in motivational enhancement: A systematic review. *Journal of Substance Abuse Treatment*, 31(1):67–73.

Mallinckrodt, C., Molenberghs, G., and Rathmann, S. (2017). Choosing estimands in clinical trials with missing data. *Pharmaceutical Statistics*, 16(1):29–36.

Maracy, M. and Dunn, G. (2011). Estimating dose-response effects in psychological treatment trials: The role of instrumental variables. *Statistical Methods in Medical Research*, 20(3):191–215.

Marshall, M., Lockwood, A., Green, G., Zajac-Roles, G., Roberts, C., and Harrison, G. (2004). Systematic assessments of need and care planning in severe mental illness: Cluster randomised controlled trial. *The British Journal of Psychiatry*, 185(2):163–168.

McLaughlin, T. J., Aupont, O., Bambauer, K. Z., Stone, P., Mullan, M. G., Colagiovanni, J., Polishuk, E., Johnstone, M., and Locke, S. E. (2005). Improving psychologic adjustment to chronic illness in cardiac patients. *Journal of General Internal Medicine*, 20(12):1084–1090.

Melis, R., Van Eijken, M., Teerenstra, S., Van Achterberg, T., Parker, S., Borm, G., Van de Lisdonk, E., Wensing, M., and Rikkert, M. O. (2008). A randomised study of a multidisciplinary programme to intervene on geriatric syndromes in vulnerable older people who live at home (Dutch EASYcare Study). *Journal of Gerontology: Medical Sciences*, 63(3):283–290.

Melmed, S., Polonsky, K. S., Larsen, P. R., and Kronenberg, H. M. (2016). *Williams textbook of endocrinology*. Elsevier Health Sciences.

Merritt, R. K., Price, J. R., Mollison, J., and Geddes, J. R. (2007). A cluster randomized controlled trial to assess the effectiveness of an intervention to educate students about depression. *Psychological Medicine*, 37(3):363–372.

Mertens, V.-C., Forsberg, L., Verbunt, J. A., Smeets, R. E., and Goossens, M. E. (2016). Treatment fidelity of a nurse-led motivational interviewing-based pre-treatment in pain rehabilitation. *The Journal of Behavioral Health Services & Research*, 43(3):459–473.

Miller, W. R. and Rollnick, S. (2002). *Motivational interviewing: Preparing people for change*. Guildford Press.

Moadel, A. B., Bernstein, S. L., Mermelstein, R. J., Arnsten, J. H., Dolce, E. H., and Shuter, J. (2012). A randomized controlled trial of a tailored group smoking cessation intervention for HIV-infected smokers. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 61(2):208–215.

Mohr, D. C., Carmody, T., Erickson, L., Jin, L., and Leader, J. (2011). Telephone-administered cognitive behavioral therapy for veterans served by community-based outpatient clinics. *Journal of Consulting & Clinical Psychology*, 79(2):261–265.

Mooney, C. Z. and Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Sage.

Moyers, T., Martin, T., Manuel, J., Miller, W., and Ernst, D. (2010). Revised global scales: Motivational interviewing treatment integrity 3.1.1 (MITI 3.1.1). *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*.

Moyers, T. B., Martin, T., Manuel, J. K., Hendrickson, S. M., and Miller, W. R. (2005). Assessing competence in the use of motivational interviewing. *Journal of Substance Abuse Treatment*, 28(1):19–26.

Moyers, T. B. and Miller, W. R. (2013). Is low therapist empathy toxic? *Psychology of Addictive Behaviors*, 27(3):878.

Muthén, L. and Muthén, B. (2012). *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén, 7th edition.

Nagelkerke, N., Fidler, V., Bernsen, R., and Borgdorff, M. (2000). Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine*, 19(14):1849–1864.

National Institute for Health and Clinical Excellence (2002). *Type 2 diabetes in adults: Management*. NICE Guidelines.

National Institute for Health and Clinical Excellence (2017). *Type 2 diabetes in adults: Management*. NICE Guidelines. Available from `www.nice.org.uk/guidance/ng28/chapter/Update-information`.

Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. (Translated and edited by DM Dabrowska and TP Speed, Statistical Science (1990), 5, 465-480). *Annals of Agricultural Sciences*, 10:1–51.

Nose, M., Barbui, C., and Tansella, M. (2003). How often do patients with psychosis fail to adhere to treatment programmes? A systematic review. *Psychological Medicine*, 33(7):1149–1160.

Onghena, P. (2005). Resentful demoralization. *Encyclopedia of Statistics in Behavioral Science*.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Perkins, K. A., Marcus, M. D., Levine, M. D., D'Amico, D., Miller, A., Broge, M., Ashcom, J., and Shiffman, S. (2001). Cognitive-behavioral therapy to reduce weight concerns improves smoking cessation outcome in weight-concerned women. *Journal of Consulting & Clinical Psychology*, 69(4):604–613.

Peyrot, M., Rubin, R. R., Lauritzen, T., Snoek, F. J., Matthews, D. R., and Skovlund, S. E. (2005). Psychosocial problems and barriers to improved diabetes management: Results of the cross-national diabetes attitudes, wishes and needs (DAWN) study. *Diabetic Medicine*, 22(10):1379–1385.

Pfiffner, L. J., Yee Mikami, A., Huang-Pollock, C., Easterlin, B., Zalecki, C., and McBurnett, K. (2007). A randomized, controlled trial of integrated home-school behavioral treatment for ADHD, predominantly inattentive type. *Journal of the American Academy of Child & Adolescent Psychiatry*, 46(8):1041–1050.

Phillips, G., Bottomley, C., Schmidt, E., Tobi, P., Lais, S., Yu, G., Lynch, R., Lock, K., Draper, A., Moore, D., Clow, A., Petticrew, M., Hayes, R., and Renton, A. (2014). Well London phase-1: Results among adults of a cluster-randomised trial of a community engagement approach to improving health behaviours and mental well-being in deprived inner-city neighbourhoods. *Journal of Epidemiology & Community Health*, 68(7):606–614.

Pickles, A. and Croudace, T. (2010). Latent mixture models for multivariate and longitudinal outcomes. *Statistical Methods in Medical Research*, 19(3):271–289.

Pinsky, P. F., Black, A., Kramer, B. S., Miller, A., Prorok, P. C., and Berg, C. (2010). Assessing contamination and compliance in the prostate component of the prostate, lung, colorectal, and ovarian (PLCO) cancer screening trial. *Clinical Trials*, 7(4):303–311.

Pocock, S. J. (2013). *Clinical trials: A practical approach*. John Wiley & Sons.

Puffer, S., Torgerson, D., Watson, J., et al. (2003). Evidence for risk of bias in cluster randomised trials: Review of recent trials published in three general medical journals. *BMJ*, 327(7418):785–789.

Radhakrishnan, M., Hammond, G., Jones, P. B., Watson, A., McMillan-Shields, F., and Lafortune, L. (2013). Cost of improving access to psychological therapies (IAPT) programme: An analysis of cost of session, treatment and recovery in selected primary care trusts in the East of England region. *Behaviour Research and Therapy*, 51(1):37–45.

Rakovshik, S. G. and McManus, F. (2010). Establishing evidence-based training in cognitive behavioral therapy: A review of current empirical findings and theoretical guidance. *Clinical Psychology Review*, 30(5):496–516.

Revicki, D. A. and Frank, L. (1999). Pharmacoeconomic evaluation in the real world. *Pharmacoeconomics*, 15(5):423–434.

Richards, D. A., Lovell, K., Gilbody, S., Gask, L., Torgerson, D., Barkham, M., Bland, M., Bower, P., Lankshear, A. J., Simpson, A., Fletcher, J., Escott, D., Hennessy, S., and Richardson, R. (2008). Collaborative care for depression in UK primary care: A randomized controlled trial. *Psychological Medicine*, 38(2):279–287.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics - Theory and Methods*, 23(8):2379–2412.

Roobol, M. J., Kerkhof, M., Schröder, F. H., Cuzick, J., Sasieni, P., Hakama, M., Stenman, U. H., Ciatto, S., Nelen, V., Kwiatkowski, M., et al. (2009). Prostate cancer mortality reduction by prostate-specific antigen–based screening adjusted for nonattendance and contamination in the european randomised study of screening for prostate cancer (ERSPC). *European Urology*, 56(4):584–591.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, volume 1, pages 20–34. American Statistical Association.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331.

Rubin, D. B. and Little, R. J. (2002). Statistical analysis with missing data. *Hoboken, NJ: J Wiley & Sons*.

Saitz, R., Cheng, D., Winter, M., Kim, T. W., Meli, S. M., Allensworth-Davies, D., Lloyd-Travaglini, C. A., and Samet, J. H. (2013). Chronic care management for dependence on alcohol and other drugs: The AHEAD randomized trial. *JAMA*, 310(11):1156–1167.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Sedgwick, P. (2014). Explanatory trials versus pragmatic trials. *BMJ*, 349:g6694.

Shaya, F. T., Yan, X., Lin, P.-J., Simoni-Wastila, L., Bron, M., Baran, R., and Donner, T. W. (2010). US trends in glycemic control, treatment, and comorbidity burden in patients with diabetes. *The Journal of Clinical Hypertension*, 12(10):826–832.

Shemilt, I., Harvey, I., Shepstone, L., Swift, L., Reading, R., Mugford, M., Belderson, P., Norris, N., Thoburn, J., and Robinson, J. (2004). A national evaluation of school breakfast clubs: Evidence from a cluster randomized controlled trial and an observational analysis. *Child: Care, Health & Development*, 30(5):413–427.

Simon, P. and Ward, N. L. (2014). An evaluation of training for lay providers in the use of motivational interviewing to promote academic achievement among urban youth. *Advances in School Mental Health Promotion*, 7(4):255–270.

Slade, M., Bird, V., Clarke, E., Le Boutillier, C., McCrone, P., Macpherson, R., Pesola, F., Wallace, G., Williams, J., and Leamy, M. (2015). Supporting recovery in patients with psychosis through care by community-based adult mental health teams (REFOCUS): A multisite, cluster, randomised, controlled trial. *The Lancet Psychiatry*, 2(6):503–514.

Slymen, D. J. and Hovell, M. F. (1997). Cluster versus individual randomization in adolescent tobacco and alcohol studies: Illustrations for design decisions. *International Journal of Epidemiology*, 26(4):765–771.

StataCorp (2015). *Stata Statistical Software: Release 14*. College Station, TX: StataCorp LP.

Stewart-Brown, S., Patterson, J., Mockford, C., Barlow, J., Klimes, I., and Pyper, C. (2004). Impact of a general practice based group parenting programme: Quantitative and qualitative results from a controlled trial at 12 months. *Archives of Disease in Childhood*, 89(6):519–525.

Stuifbergen, A. K., Blozis, S. A., Becker, H., Phillips, L., Timmerman, G., Kullberg, V., Taxis, C., and Morrison, J. (2010). A randomized controlled trial of a wellness intervention for women with fibromyalgia syndrome. *Clinical Rehabilitation*, 24(4):305–318.

Taylor, J. L., Novaco, R. W., Gillmer, B. T., Robertson, A., and Thorne, I. (2005). Individual cognitive-behavioural anger treatment for people with mild-borderline intellectual disabilities and histories of aggression: A controlled trial. *British Journal of Clinical Psychology*, 44(3):367–382.

Tchetgen, E. J. T. and VanderWeele, T. J. (2012). On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75.

Tchetgen, E. J. T., Walter, S., Vansteelandt, S., Martinussen, T., and Glymour, M. (2015). Instrumental variable estimation in a survival context. *Epidemiology*, 26(3):402.

Teerenstra, S., Melis, R., Peer, P., and Borm, G. (2006). Pseudo cluster randomization dealt with selection bias and contamination in clinical trials. *Journal of Clinical Epidemiology*, 59(4):381–386.

Terza, J. V., Basu, A., and Rathouz, P. J. (2008). Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics*, 27(3):531–543.

Thorn, B. E., Day, M. A., Burns, J., Kuhajda, M. C., Gaskins, S. W., Sweeney, K., McConley, R., Ward, C. L., and Cabbil, C. (2011). Randomized trial of group cognitive behavioral therapy compared with a pain education control for low-literacy rural people with chronic pain. *Pain*, 152(12):2710–2720.

Tiwari, A., Leung, W. C., Leung, T. W., Humphreys, J., Parker, B., and Ho, P. C. (2005). A randomised controlled trial of empowerment training for Chinese abused pregnant women in Hong Kong. *BJOG: An International Journal of Obstetrics & Gynaecology*, 112(9):1249–1256.

Torgerson, D. J. (2001). Contamination in trials: Is cluster randomisation the answer? *BMJ*, 322(7282):355–357.

Van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18(6):681–694.

VanderWeele, T. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press.

Vanderweele, T. J., Hong, G., Jones, S. M., and Brown, J. L. (2013). Mediation and spillover effects in group-randomized trials: A case study of the 4Rs educational intervention. *Journal of the American Statistical Association*, 108(502):469–482.

Waghorn, G., Dias, S., Gladman, B., Harris, M., and Saha, S. (2014). A multi-site randomised controlled trial of evidence-based supported employment for adults with severe and persistent mental illness. *Australian Occupational Therapy Journal*, 61(6):424–436.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics*, 11(3):284–300.

Walpole, B., Dettmer, E., Morrongiello, B. A., McCrindle, B. W., and Hamilton, J. (2013). Motivational interviewing to enhance self-efficacy and promote weight loss in over-

weight and obese adolescents: A randomized controlled trial. *Journal of Pediatric Psychology*, 38(9):944–953.

Watkins, C., Huang, X., Latimer, N., Tang, Y., and Wright, E. J. (2013). Adjusting overall survival for treatment switches: Commonly used methods and practical application. *Pharmaceutical Statistics*, 12(6):348–357.

Weaver, T., Metrebian, N., Hellier, J., Pilling, S., Charles, V., Little, N., Poovendran, D., Mitcheson, L., Ryan, F., Bowden-Jones, O., Dunn, J., Glasper, A., Finch, E., and Strang, J. (2014). Use of contingency management incentives to improve completion of hepatitis B vaccination in people undergoing treatment for heroin dependence: A cluster randomised trial. *The Lancet*, 384(9938):153–163.

Wells, K. B., Sherbourne, C., Schoenbaum, M., Duan, N., Meredith, L., Unützer, J., Miranda, J., Carney, M. F., and Rubenstein, L. V. (2000). Impact of disseminating quality improvement programs for depression in managed primary care: A randomized controlled trial. *JAMA*, 283(2):212–220.

Welsh, A. (2013). Randomised controlled trials and clinical maternity care: Moving on from intention-to-treat and other simplistic analyses of efficacy. *BMC Pregnancy and Childbirth*, 13(15).

Wennberg, J. E., Barry, M. J., Fowler, F. J., and Mulley, A. (1993). Outcomes research, ports, and health care reform. *Annals of the New York Academy of Sciences*, 703(1):52–62.

White, H. (1984). *Asymptotic theory for econometricians*. Academic press.

White, I. R. (2005). Uses and limitations of randomization-based efficacy estimators. *Statistical Methods in Medical Research*, 14(4):327–347.

Whittemore, R., Melkus, G., Sullivan, A., and Grey, M. (2003). A nurse-coaching intervention for women with type 2 diabetes. *The Diabetes Educator*, 30(5):795–804.

Whittle, A., Hettema, J., Manuel, J., Cangelosi, C., Coffa, D., De La Cerda, S., Tierney, M., and Lum, P. (2015). Effectiveness of continuing education in motivational interviewing for health professionals working with families and pediatric patients: Results of a skills-based assessment. *Family Medicine & Medical Science Research*, 4(187):2.

Wood, J. J., Piacentini, J. C., Southam-Gerow, M., Chu, B. C., and Sigman, M. (2006). Family cognitive behavioral therapy for child anxiety disorders. *Journal of the American Academy of Child & Adolescent Psychiatry*, 45(3):314–321.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources*, 50(2):420–445.

Zelen, M. (1979). A new design for randomized clinical trials. *New England Journal of Medicine*, 300(22):1242–1245.

Appendix A

Scoping review

The sections of this appendix refer to the scoping review of problems and solutions associated with contamination in mental health trials. This was the subject of Chapter 2.

A.1 Ovid search procedure

The steps used in the Ovid search procedure for trials of complex interventions in mental health where contamination was a problem are listed below.

1. (randomized controlled trial or randomised controlled trial).pt,dt.
2. controlled clinical trial.pt,dt.
3. (randomized or randomised).ab.
4. placebo.ab.
5. clinical trial.sh.
6. randomly.ab.
7. trial.ti.
8. 1 or 2 or 3 or 4 or 5 or 6 or 7
9. (animals not humans).sh.
10. 8 not 9
11. (complex adj3 intervention).af.
12. (complex adj3 treatment).af.
13. (complex adj3 training).af.
14. (multicomponent adj3 intervention).af.
15. (multicomponent adj3 treatment).af.
16. (multicomponent adj3 training).af.

17. (multifaceted adj3 intervention).af.
18. (multifaceted adj3 treatment).af.
19. (multifaceted adj3 training).af.
20. (social adj3 intervention).af.
21. (social adj3 treatment).af.
22. (social adj3 training).af.
23. (psychological adj3 intervention).af.
24. (psychological adj3 treatment).af.
25. (psychological adj3 training).af.
26. (psychological adj3 therapy).af.
27. (psychosocial adj3 intervention).af.
28. (psychosocial adj3 treatment).af.
29. (psychosocial adj3 training).af.
30. psychotherap*.af.
31. therapist.af.
32. (behavio?ral adj3 intervention).af.
33. (behavio?ral adj3 treatment).af.
34. (behavio?ral adj3 training).af.
35. 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34
36. (treatment adj6 contaminat*).af.
37. (arm adj6 contaminat*).af.
38. (control adj6 contaminat*).af.
39. (group*1 adj6 contaminat*).af.
40. (outcome adj6 contaminat*).af.
41. (trial adj6 contaminat*).af.
42. (patient*1 adj6 contaminat*).af.
43. (intervention adj6 contaminat*).af.
44. (treatment adj6 spillover).af.
45. (arm adj6 spillover).af.
46. (control adj6 spillover).af.
47. (group*1 adj6 spillover).af.
48. (outcome adj6 spillover).af.
49. (trial adj6 spillover).af.

50. (patient*1 adj6 spillover).af.
51. (intervention adj6 spillover).af.
52. 36 or 37 or 38 or 39 or 40 or 41 or 42 or 43 or 44 or 45 or 46 or 47 or 48 or 49 or 50 or 51
53. (blood adj3 contaminat*).af.
54. (microb* adj3 contaminat*).af.
55. 53 or 54
56. 52 not 55
57. 10 and 35 and 56
58. device.ti.
59. device.sh.
60. device.ab.
61. vaccine.ti.
62. vaccine.sh.
63. vaccine.ab.
64. surgery.ti.
65. surgery.sh.
66. surgery.ab.
67. microb*.ti.
68. microb*.sh.
69. microb*.ab.
70. antimicrob*.ti.
71. antimicrob*.sh.
72. antimicrob*.ab.
73. (genes or genetic*).ti.
74. (genes or genetic*).sh.
75. (genes or genetic*).ab.
76. screening.ti.
77. decision aid.ti.
78. decision support.ti.
79. 58 or 59 or 60 or 61 or 62 or 63 or 64 or 65 or 66 or 67 or 68 or 69 or 70 or 71 or 72 or 73 or 74 or 75 or 76 or 77 or 78
80. 57 not 79
81. limit 80 to yr=“2000 -Current”

82. mental health.af.
83. psycholog*.af.
84. psychiatr*.af.
85. 82 or 83 or 84
86. 81 and 85

A.2 Pro forma for information/data extraction

Each article that was found using the search procedure above was assessed using five inclusion criteria. Articles that met all five criteria were then assessed for bias, trial design, contamination process and particular parameters of interest. The inclusion criteria and assessment items are listed below.

Inclusion criteria

Is it within the context of a specific randomised controlled clinical trial? (Yes; no)
Is it a complex intervention? I.e. an intervention where multiple components work together to produce some extra benefit. (Yes; no)
Does it discuss the process that lead to contamination of the control arm? (Yes; no)
Is it in English? (Yes; no)
Is it mental health related (target population; intervention; primary outcome)? (Yes; no)

Assessment of bias

Jadad score (0-5)
Was allocation sequence adequately generated? (Selection bias) (Done; not done; unclear)
Was allocation adequately concealed? (Selection bias) (Done; not done; unclear)
Did randomisation occur after consent was obtained? (Selection bias) (Done; not done; unclear)
Were baseline measures completed before randomisation? (Selection bias) (Done; not done; unclear)
Were baseline outcome measurements similar across trial arms? (Selection bias) (Done; not done; unclear)
Were baseline characteristics similar across trial arms? (Selection bias) (Done; not done; unclear)

Was knowledge of allocation adequately prevented during study? (Performance bias) (Done; not done; unclear)
Was assessment blinded? (Assessor bias) (Done; not done; unclear)
Were incomplete outcome data adequately addressed? (Attrition bias) (Done; not done; unclear)
Was there evidence of similar attrition in the trial arms? (Attrition bias) (Done; not done; unclear)
Was study free from selective outcome reporting? (Reporting bias) (Done; not done; unclear)
Was study free from other risks of bias? (Other biases) (Done; not done; unclear)
If it is a late phase trial, is a sample size calculation report? (Done; not done; unclear)
If it is a cluster RCT (and late phase), was sample size inflated for clustering? (Done; not done; unclear)

If qualifying, trial design

Type of trial (early phase; late phase)
What is the target population? (Free text)
Definition of intervention. (Free text)
Definition of comparison group. (Free text)
Number of trial arms. (Numeric)
What is the primary outcome? (Free text)
Length of follow-up (months). (Free text)
Unit of randomisation. (Individual; cluster)
If cluster randomisation, what is the cluster? (Free text)
At what level is the intervention provided? (Participant; clinician; school class; organisation; community)

Contamination (describing contamination process)

How was contamination anticipated or thought to have taken place? (free text)
Was it considered to have resulted from communication? (yes; no)
If yes, was communication thought to occur at the same level as the intervention was provided? (yes; no)
When was contamination anticipated / observed to take place? (Any time after intervention; During delivery of intervention; During data collection; Unclear)

Was there empirical evidence that was thought to represent the effect of contamination (or lack of)? (yes; no)

If yes, what was this evidence? (free text)

Were any actions (apart from cluster randomisation) proposed to address contamination? (yes; no)

If yes, what were these? (free text)

Was cluster randomisation used to avoid contamination? (yes; no)

If yes, what was the rationale for the choice of the cluster? (free text)

Simulation parameters

From sample size calculation:

- Predicted average cluster size. (Numeric)
- Predicted range of cluster sizes
- Predicted intraclass correlation coefficient
- Predicted design effect (1+ICC(K-1))
- Planned power (for main analysis)
- Alpha level
- One or two-sided test. (One-sided; two-sided)
- Anticipated drop-out
- Sample size intention
- Anticipated treatment effect. (Numeric)
- Type of effect (Cohen's d; difference in proportions; difference in means; odds ratio; hazard ratio)
- Standardised treatment effect for power calculation

Achieved sample size

Size of control group

Observed ICC for cluster randomised/group intervention trials

Observed mean cluster size for cluster randomised trials

Treatment effect (numeric)

Type of effect (estimated/observed Cohen's d; estimated/observed difference in proportions; estimated/observed odds ratio; estimated/observed hazard ratio; estimated/observed rate ratio)

Any non-relevant fields were marked "Not given"; any unaddressed fields were marked "Not given"; anything that was not described clearly enough was marked "Unclear".

A.3 Articles from which data were abstracted in the scoping review

References for the 234 articles that constituted the final set of the scoping review are listed below.

[1] Aiarzaguena JM, Grandes G, Gaminde I, Salazar A, Sanchez A, Arino J. A randomized controlled clinical trial of a psychosocial and communication intervention carried out by GPs for patients with medically unexplained symptoms. Psychological Medicine. 2007;37:283-94.

[2] Albers L, Heinen F, Landgraf M, Straube A, Blum B, Filippopulos F, et al. Headache cessation by an educational intervention in grammar schools: a cluster randomized trial. European Journal of Neurology. 2015;22:270-e22.

[3] Alessi CA, Martin JL, Webber AP, Kim CE, Harker JO, Josephson KR. Randomized, Controlled Trial of a Nonpharmacological Intervention to Improve Abnormal Sleep/-Wake Patterns in Nursing Home Residents. Journal of the American Geriatrics Society. 2005;53:803-10.

[4] Alexopoulos GS, Reynolds CF, Bruce ML, Katz IR, Raue PJ, Mulsant BH, et al. Reducing Suicidal Ideation and Depression in Older Primary Care Patients: 24-Month Outcomes of the PROSPECT Study. American Journal of Psychiatry. 2009;166:882-90.

[5] Anderson BK, Larimer ME. Problem Drinking and the Workplace: An Individualized Approach to Prevention. Psychology of Addictive Behaviors. 2002;16:243-51.

[6] Araya R, Flynn T, Rojas G, Fritsch R, Simon G. Cost-Effectiveness of a Primary Care Treatment Program for Depression in Low-Income Women in Santiago, Chile. American Journal of Psychiatry. 2006;163:1379-87.

[7] Arean PA, Gum A, McCulloch CE, Bostrom A, Gallagher-Thompson D, Thompson L. Treatment of Depression in Low-Income Older Adults. Psychology & Aging. 2005;20:601-9.

[8] Atlantis E, Chow C-M, Kirby A, Fiatarone Singh MA. Worksite Intervention Effects on Sleep Quality: A Randomized Controlled Trial. Journal of Occupational Health Psychology. 2006;11:291-304.

[9] Aune T, Stiles TC. Universal-Based Prevention of Syndromal and Subsyndromal Social Anxiety: A Randomized Controlled Study. Journal of Consulting & Clinical Psychology. 2009;77:867-79.

[10] Aveyard P, Brown K, Saunders C, Alexander A, Johnstone E, Munafo MR, et al. Weekly versus basic smoking cessation support in primary care: a randomised controlled trial. Thorax. 2007;62:898-903.

[11] Baillargeon L, Landreville P, Verreault R, Beauchemin J-P, Gregoire J-P, Morin CM. Discontinuation of benzodiazepines among older insomniac adults treated with cognitive-behavioural therapy combined with gradual tapering: a randomized trial. CMAJ Canadian Medical Association Journal. 2003;169:1015-20.

[12] Baker-Henningham H, Scott S, Jones K, Walker S. Reducing child conduct problems and promoting social skills in a middle-income country: cluster randomised controlled trial +. British Journal of Psychiatry. 2012;201:101-8.

[13] Ball SA, Martino S, Nich C, Frankforter TL, Van Horn D, Crits-Christoph P, et al. Site Matters: Multisite Randomized Trial of Motivational Enhancement Therapy in Community Drug Abuse Clinics. Journal of Consulting & Clinical Psychology. 2007;75:556-67.

[14] Bannink R, Broeren S, Heydelberg J, van 't Klooster E, van Baar C, Raat H. Your Health, an intervention at senior vocational schools to promote adolescents' health and health behaviors. Health Education Research. 2014;29:773-85.

[15] Barkhof E, Meijer CJ, de Sonneville LMJ, Linszen DH, de Haan L. The Effect of Motivational Interviewing on Medication Adherence and Hospitalization Rates in Nonadherent Patients with Multi-Episode Schizophrenia. Schizophrenia Bulletin. 2013;39:1242-51.

[16] Barsky AJ, Ahern DK. Cognitive Behavior Therapy for Hypochondriasis: A Randomized Controlled Trial. JAMA. 2004;291:1464-70.

[17] Barton MB, Morley DS, Moore S, Allen JD, Kleinman KP, Emmons KM, et al. Decreasing Women's Anxieties After Abnormal Mammograms: A Controlled Trial. Journal of the National Cancer Institute. 2004;96:529-38.

[18] Bazargan-Hejazi S, Bing E, Bazargan M, Der-Martirosian C, Hardin E, Bernstein J, et al. Evaluation of a Brief Intervention in an Inner-City Emergency Department. Annals of Emergency Medicine. 2005;46:67-76.

[19] Beach SRH, Kogan SM, Brody GH, Chen Y-F, Lei M-K, Murry VM. Change in Caregiver Depression as a Function of the Strong African American Families Program. Journal of Family Psychology. 2008;22:241-52.

[20] Beaver K, Campbell M, Williamson S, Procter D, Sheridan J, Heath J, et al. An exploratory randomized controlled trial comparing telephone and hospital follow-up after treatment for colorectal cancer. Colorectal Disease. 2012;14:1201-9.

[21] Beck CK, Vogelpohl TS, Rasin JH, Uriri JT, O'Sullivan P, Walls R, et al. Effects of

Behavioral Interventions on Disruptive Behavior and Affect in Demented Nursing Home Residents. Nursing Research July/August. 2002;51:219-28.
[22] Beckie TM, Beckstead JW. Predicting Cardiac Rehabilitation Attendance in a Gender-Tailored Randomized Clinical Trial. Journal of Cardiopulmonary Rehabilitation & Prevention May/June. 2010;30:147-56.
[23] Becona E, Vazquez FL. Effectiveness of Personalized Written Feedback Through a Mail Intervention for Smoking Cessation: A Randomized-Controlled Trial in Spanish Smokers. Journal of Consulting & Clinical Psychology. 2001;69:33-40.
[24] Befort CA, Nollen N, Ellerbeck EF, Sullivan DK, Thomas JL, Ahluwalia JS. Motivational interviewing fails to improve outcomes of a behavioral weight loss program for obese African American women: a pilot randomized trial. Journal of Behavioral Medicine. 2008;31:367-77.
[25] Bellantonio S, Kenny AM, Fortinsky RH, Kleppinger A, Robison J, Gruman C, et al. Efficacy of a Geriatrics Team Intervention for Residents in Dementia-Specific Assisted Living Facilities: Effect on Unanticipated Transitions. Journal of the American Geriatrics Society. 2008;56:523-8.
[26] Bermejo I, Schneider F, Kriston L, Gaebel W, Hegerl U, Berger M, et al. Improving outpatient care of depression by implementing practice guidelines: a controlled clinical trial. International Journal for Quality in Health Care. 2009;21:29-36.
[27] Bernstein GA, Layne AE, Egan EA, Tennison DM. School-Based Interventions for Anxious Children. Journal of the American Academy of Child & Adolescent Psychiatry. 2005;44:1118-27.
[28] Bernstein SL, Bijur P, Cooperman N, Jearld S, Arnsten JH, Moadel A, et al. A Randomized Trial of a Multicomponent Cessation Strategy for Emergency Department Smokers. Academic Emergency Medicine. 2011;18:575-83.
[29] Bombardier CH, Bell KR, Temkin NR, Fann JR, Hoffman J, Dikmen S. The Efficacy of a Scheduled Telephone Intervention for Ameliorating Depressive Symptoms During the First Year After Traumatic Brain Injury. Journal of Head Trauma Rehabilitation July/August. 2009;24:230-8.
[30] Borland R, Balmford J, Benda P. Population-level effects of automated smoking cessation help programs: a randomized controlled trial. Addiction. 2013;108:618-28.
[31] Bormann JE, Gifford AL, Shively M, Smith TL, Redwine L, Kelly A, et al. Effects of Spiritual Mantram Repetition on HIV Outcomes: A Randomized Controlled Trial. Journal of Behavioral Medicine. 2006;29:359-76.

[32] Bosmans JE, Brook OH, van Hout HPJ, de Bruijne MC, Nieuwenhuyse H, Bouter LM, et al. Cost Effectiveness of a Pharmacy-Based Coaching Programme to Improve Adherence to Antidepressants. Pharmacoeconomics. 2007;25:25-37.

[33] Bricker JB, Bush T, Zbikowski S, Mercer LD, Heffner JL. Randomized Trial of Telephone-Delivered Acceptance and Commitment Therapy Versus Cognitive Behavioral Therapy for Smoking Cessation: A Pilot Study. Nicotine & Tobacco Research. 2014;16:1446-54.

[34] Brouwers EPM, de Bruijne MC, Terluin B, Tiemens BG, Verhaak PFM. Cost-effectiveness of an activating intervention by social workers for patients with minor mental disorders on sick leave: a randomized controlled trial. European Journal of Public Health. 2007;17:214-20.

[35] Brown RA, Ramsey SE, Strong DR, Myers MG, Kahler CW, Lejuez CW, et al. Effects of motivational interviewing on smoking cessation in adolescents with psychiatric disorders. Tobacco Control. 2003;12 Supplement:iv3-iv10.

[36] Bruce ML, Ten Have TR, Reynolds CF, Katz II, Schulberg HC, Mulsant BH, et al. Reducing Suicidal Ideation and Depressive Symptoms in Depressed Older Primary Care Patients: A Randomized Controlled Trial. JAMA. 2004;291:1081-91.

[37] Burling TA, Burling AS, Latini D. A Controlled Smoking Cessation Trial for Substance-Dependent Inpatients. Journal of Consulting & Clinical Psychology. 2001;69:295-304.

[38] Calear AL, Christensen H, Mackinnon A, Griffiths KM, O'Kearney R. The YouthMood Project: A Cluster Randomized Controlled Trial of an Online Cognitive Behavioral Program With Adolescents. Journal of Consulting & Clinical Psychology. 2009;77:1021-32.

[39] Callahan CM, Boustani MA, Unverzagt FW, Austrom MG, Damush TM, Perkins AJ, et al. Effectiveness of Collaborative Care for Older Adults With Alzheimer Disease in Primary Care: A Randomized Controlled Trial. JAMA. 2006;295:2148-57.

[40] Campbell-Heider N, Tuttle J, Knapp TR. The Effect of Positive Adolescent Life Skills Training on Long Term Outcomes for High-Risk Teens. Journal of Addictions Nursing. 2009;20:6-15.

[41] Cappella E, Hamre BK, Kim HY, Henry DB, Frazier SL, Atkins MS, et al. Teacher Consultation and Coaching Within Mental Health Practice: Classroom and Child Effects in Urban Elementary Schools. Journal of Consulting & Clinical Psychology. 2012;80:597-610.

[42] Chan MF, Ng SE, Tien A, Man Ho RC, Thayala J. A randomised controlled study to explore the effect of life story review on depression in older Chinese in Singapore.

Health & Social Care in the Community. 2013;21:545-53.
[43] Chanen AM, Jackson HJ, McCutcheon LK, Jovev M, Dudgeon P, Yuen HP, et al. Early intervention for adolescents with borderline personality disorder using cognitive analytic therapy: randomised controlled trial. British Journal of Psychiatry. 2008;193:477-84.
[44] Chang M-Y, Chen C-H, Huang K-F. Effects of music therapy on psychological health of women during pregnancy. Journal of Clinical Nursing. 2008;17:2580-7.
[45] Chen Z, Meng Z, Milbury K, Bei W, Zhang Y, Thornton B, et al. Qigong improves quality of life in women undergoing radiotherapy for breast cancer: Results of a randomized controlled trial. Cancer. 2013;119:1690-8.
[46] Chochinov HM, Kristjanson LJ, Breitbart W, McClement S, Hack TF, Hassard T, et al. Effect of dignity therapy on distress and end-of-life experience in terminally ill patients: a randomised controlled trial. Lancet Oncology. 2011;12:753-62.
[47] Chouinard M-C, Robichaud-Ekstrand S. The Effectiveness of a Nursing Inpatient Smoking Cessation Program in Individuals With Cardiovascular Disease. Nursing Research July/August. 2005;54:243-54.
[48] Clarke AM, Bunting B, Barry MM. Evaluating the implementation of a school-based emotional well-being programme: a cluster randomized controlled trial of Zippy's Friends for children in disadvantaged primary schools. Health Education Research. 2014;29:786-98.
[49] Clarkson JE, Young L, Ramsay CR, Bonner BC, Bonetti D. How to influence patient oral hygiene behavior effectively. Journal of Dental Research. 2009;88:933-7.
[50] Clement S, van Nieuwenhuizen A, Kassam A, Flach C, Lazarus A, de Castro M, et al. Filmed v. live social contact interventions to reduce stigma: randomised controlled trial. British Journal of Psychiatry. 2012;201:57-64.
[51] Cole MG, McCusker J, Bellavance F, Primeau FJ, Bailey RF, Bonnycastle MJ, et al. Systematic detection and multidisciplinary care of delirium in older medical inpatients: a randomized trial. CMAJ Canadian Medical Association Journal. 2002;167:753-9.
[52] Cole MG, McCusker J, Elie M, Dendukuri N, Latimer E, Belzile E. Systematic detection and multidisciplinary care of depression in older medical inpatients: a randomized trial. CMAJ Canadian Medical Association Journal. 2006;174:38-44.
[53] Connors GJ, Walitzer KS, Dermen KH. Preparing Clients for Alcoholism Treatment: Effects on Treatment Participation and Outcomes. Journal of Consulting & Clinical Psychology. 2002;70:1161-9.
[54] Cook S, Chambers E, Coleman JH. Occupational therapy for people with psychotic

conditions in community settings: a pilot randomized controlled trial. Clinical Rehabilitation. 2009;23:40-52.

[55] Cooper LA, Ghods Dinoso BK, Ford DE, Roter DL, Primm AB, Larson SM, et al. Comparative Effectiveness of Standard versus Patient-Centered Collaborative Care Interventions for Depression among African Americans in Primary Care Settings: The BRIDGE Study. Health Services Research. 2013;48:150-74.

[56] Copello A, Templeton L, Orford J, Velleman R, Patel A, Moore L, et al. The relative efficacy of two levels of a primary care intervention for family members affected by the addiction problem of a close relative: a randomized trial. Addiction. 2009;104:49-58.

[57] Courneya KS, Friedenreich CM, Sela RA, Quinney H, Rhodes RE, Handman M. The group psychotherapy and home-based physical exercise (group-hope) trial in cancer survivors: Physical fitness and quality of life outcomes. Psycho-Oncology. 2003;12:357-74.

[58] Coventry P, Lovell K, Dickens C, Bower P, Chew-Graham C, McElvenny D, et al. Integrated primary care for patients with mental and physical multimorbidity: cluster randomised controlled trial of collaborative care for patients with depression comorbid with diabetes or cardiovascular disease. BMJ February. 2015;14.

[59] Craig TKJ, Johnson S, McCrone P, Afuwape S, Hughes E, Gournay K, et al. Integrated Care for Co-occurring Disorders: Psychiatric Symptoms, Social Functioning, and Service Costs at 18 Months. Psychiatric Services. 2008;59:276-82.

[60] Cullen AE, Clarke AY, Kuipers E, Hodgins S, Dean K, Fahy T. A Multisite Randomized Trial of a Cognitive Skills Program for Male Mentally Disordered Offenders: Violence and Antisocial Behavior Outcomes. Journal of Consulting & Clinical Psychology. 2012;80:1114-20.

[61] Dakof GA, Henderson CE, Rowe CL, Boustani M, Greenbaum PE, Wang W, et al. A randomized clinical trial of family therapy in juvenile drug court. Journal of family psychology : JFP : journal of the Division of Family Psychology of the American Psychological Association. 2015;29:232-41.

[62] De Wit M, Delemarre-Van De Waal HA, Bokma JA, Haasnoot K, Houdijk MC, Gemke RJ, et al. Monitoring and Discussing Health-Related Quality of Life in Adolescents With Type 1 Diabetes Improve Psychosocial Well-Being: A randomized controlled trial. Diabetes Care. 2008;31:1521-6.

[63] Dechamps A, Alban R, Jen J, Decamps A, Traissac T, Dehail P. Individualized Cognition-Action intervention to prevent behavioral disturbances and functional de-

cline in institutionalized older adults: a randomized pilot trial. International Journal of Geriatric Psychiatry. 2010;25:850-60.

[64] Dennis M, Titus JC, Diamond G, Donaldson J, Godley SH, Tims FM, et al. The Cannabis Youth Treatment (CYT) experiment: rationale, study design and analysis plans. Addiction Supplement. 2002;97 Supplement:16-34.

[65] Deudon A, Maubourguet N, Gervais X, Leone E, Brocker P, Carcaillon L, et al. Non-pharmacological management of behavioural symptoms in nursing homes. International Journal of Geriatric Psychiatry. 2009;24:1386-95.

[66] Dilley JW, Woods WJ, Loeb L, Nelson K, Sheon N, Mullan J, et al. Brief Cognitive Counseling With HIV Testing To Reduce Sexual Risk Among Men Who Have Sex With Men: Results From a Randomized Controlled Trial Using Paraprofessional Counselors. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2007;44:569-77.

[67] Dobscha SK, Corson K, Perrin NA, Hanson GC, Leibowitz RQ, Doak MN, et al. Collaborative Care for Chronic Pain in Primary Care: A Cluster Randomized Trial. JAMA. 2009;301:1242-52.

[68] Dracup K, Moser DK, Doering LV, Guzy PM, Juarbe T. A controlled trial of cardiopulmonary resuscitation training for ethnically diverse parents of infants at high risk for cardiopulmonary arrest. Critical Care Medicine. 2000;28:3289-95.

[69] Ell K, Katon W, Xie B, Lee P-J, Kapetanovic S, Guterman J, et al. Collaborative Care Management of Major Depression Among Low-Income, Predominantly Hispanic Subjects With Diabetes: A randomized controlled trial. Diabetes Care. 2010;33:706-13.

[70] Ersek M, Turner JA, Cain KC, Kemp CA. Results of a randomized controlled trial to examine the efficacy of a chronic pain self-management group for older adults [ISRCTN11899548]. Pain. 2008;138:29-40.

[71] Farooq S, Nazar Z, Irfan M, Akhter J, Gul E, Irfan U, et al. Schizophrenia medication adherence in a resource-poor setting: randomised controlled trial of supervised treatment in out-patients for schizophrenia (STOPS). British Journal of Psychiatry. 2011;199:467-72.

[72] van der Feltz-Cornelis CM, van Oppen P, Ader HJ, van Dyck R. Randomised Controlled Trial of a Collaborative Care Model with Psychiatric Consultation for Persistent Medically Unexplained Symptoms in General Practice. Psychotherapy & Psychosomatics. 2006;75:282-9.

[73] Finnema E, Droes R-M, Ettema T, Ooms M, Ader H, Ribbe M, et al. The effect of integrated emotion-oriented care versus usual care on elderly persons with dementia in

the nursing home and on nursing assistants: a randomized clinical trial. International Journal of Geriatric Psychiatry. 2005;20:330-43.
[74] Forchuk C, Martin ML, Chan YL, Jensen E. Therapeutic relationships: from psychiatric hospital to community. Journal of Psychiatric & Mental Health Nursing. 2005;12:556-64.
[75] Foshee VA, Bauman KE, Greene WF, Koch GG, Linder GF, MacDougall JE. The Safe Dates Program: 1-Year Follow-Up Results. American Journal of Public Health Disease Elimination and Eradication. 2000;90:1619-22.
[76] Fossey J, Ballard C, Juszczak E, James I, Alder N, Jacoby R, et al. Effect of enhanced psychosocial care on antipsychotic use in nursing home residents with severe dementia: cluster randomised trial. BMJ. 2006;332:756-61.
[77] Gallagher R, McKinley S, Dracup K. Effects of a telephone counseling intervention on psychosocial adjustment in women following a cardiac event. Heart & Lung: Journal of Acute & Critical Care March/April. 2003;32:79-87.
[78] Garcia C, Pintor J, Vazquez G, Alvarez-Zumarraga E. Project Wings, a Coping Intervention for Latina Adolescents: A Pilot Study. Western Journal of Nursing Research. 2013;35:434-58.
[79] Garner BR, Godley SH, Dennis ML, Hunter BD, Bair CM, Godley MD. Using Pay for Performance to Improve Treatment Implementation for Adolescent Substance Use Disorders: Results From a Cluster Randomized Trial. Archives of Pediatrics & Adolescent Medicine. 2012;166:938-44.
[80] Gartner FR, Nieuwenhuijsen K, Ketelaar SM, van Dijk FJ, Sluiter JK. The Mental Vitality @ Work Study: Effectiveness of a Mental Module for Workers' Health Surveillance for Nurses and Allied Health Care Professionals on Their Help-Seeking Behavior. Journal of Occupational & Environmental Medicine. 2013;55:1219-29.
[81] Gater R, Waheed W, Husain N, Tomenson B, Aseem S, Creed F. Social intervention for British Pakistani women with depression: randomised controlled trial. British Journal of Psychiatry. 2010;197:227-33.
[82] Gensichen J, von Korff M, Peitz M, Muth C, Beyer M, Guthlin C, et al. Case Management for Depression by Health Care Assistants in Small Primary Care Practices: A Cluster Randomized Trial. Annals of Internal Medicine. 2009;151:369-78.
[83] Glazebrook C, Marlow N, Israel C, Croudace T, Johnson S, White IR, et al. Randomised trial of a parenting intervention during neonatal intensive care. Archives of Disease in Childhood Fetal & Neonatal Edition. 2007;92:F438-F43.
[84] de Godoy DV, de Godoy RF. A randomized controlled trial of the effect of psychot-

herapy on anxiety and depression in chronic obstructive pulmonary disease. Archives of physical medicine and rehabilitation. 2003;84:1154-7.

[85] Gold DT, Shipp KM, Pieper CF, Duncan PW, Martinez S, Lyles KW. Group Treatment Improves Trunk Strength and Psychological Status in Older Women with Vertebral Fractures: Results of a Randomized, Clinical Trial. Journal of the American Geriatrics Society. 2004;52:1471-8.

[86] Han C-K, Ssewamala FM, Wang JS-H. Family economic empowerment and mental health among AIDS-affected children living in AIDS-impacted communities: evidence from a randomised evaluation in southwestern Uganda. Journal of Epidemiology & Community Health. 2013;67:225-30.

[87] Hansson L, Svensson B, Bjorkman T, Bullenkamp J, Lauber C, Martinez-Leal R, et al. What works for whom in a computer-mediated communication intervention in community psychiatry? Moderators of outcome in a cluster randomized trial. Acta Psychiatrica Scandinavica. 2008;118:404-9.

[88] Harvey AG, Belanger L, Talbot L, Eidelman P, Beaulieu-Bonneau S, Fortier-Brochu E, et al. Comparative Efficacy of Behavior Therapy, Cognitive Therapy, and Cognitive Behavior Therapy for Chronic Insomnia: A Randomized Controlled Trial. Journal of Consulting & Clinical Psychology. 2014;82:670-83.

[89] Haukka E, Pehkonen I, Leino-Arjas P, Viikari-Juntura E, Takala E-P, Malmivaara A, et al. Effect of a participatory ergonomics intervention on psychosocial factors at work in a randomised controlled trial. Occupational & Environmental Medicine. 2010;67:170-7.

[90] Hayes L, Boyd C, Sewell J. Acceptance and commitment therapy for the treatment of adolescent depression: A pilot study in a psychiatric outpatient setting. Mindfulness. 2011;2:86-94.

[91] Heirich M, Sieck CJ. Worksite cardiovascular wellness programs as a route to substance abuse prevention. Journal of occupational and environmental medicine. 2000;42:47-56.

[92] Herz MI, Lamberti JS, Mintz J, Scott R, O'Dell SP, McCartan L, et al. A Program for Relapse Prevention in Schizophrenia: A Controlled Study. Archives of General Psychiatry. 2000;57:277-83.

[93] Hjorthoj CR, Fohlmann A, Larsen AM, Gluud C, Arendt M, Nordentoft M. Specialized psychosocial treatment plus treatment as usual (TAU) versus TAU for patients with cannabis use disorder and psychosis: the CapOpus randomized trial. Psychological Medicine. 2013;43:1499-510.

[94] Hoffman SL, Hanrahan SJ. Mental Skills for Musicians: Managing Music Performance Anxiety and Enhancing Performance. Sport, Exercise, & Performance Psychology. 2012;1:17-28.

[95] Horowitz JA, Murphy CA, Gregory K, Wojcik J, Pulcini J, Solon L. Nurse Home Visits Improve Maternal/Infant Interaction and Decrease Severity of Postpartum Depression. JOGNN Journal of Obstetric, Gynecologic, & Neonatal Nursing May/June. 2013;42:287-300.

[96] Horowitz JL, Garber J, Ciesla JA, Young JF, Mufson L. Prevention of Depressive Symptoms in Adolescents: A Randomized Trial of Cognitive-Behavioral and Interpersonal Prevention Programs. Journal of Consulting & Clinical Psychology. 2007;75:693-706.

[97] Horrell L, Goldsmith KA, Tylee AT, Schmidt UH, Murphy CL, Bonin E-M, et al. One-day cognitive-behavioural therapy self-confidence workshops for people with depression: randomised controlled trial. British Journal of Psychiatry. 2014;204:222-33.

[98] van den Hout JHC, Vlaeyen JWS, Heuts PHTG, Zijlema JHL, Wijnen JAG. Secondary Prevention of Work-Related Disability in Nonspecific Low Back Pain: Does Problem-Solving Therapy Help? A Randomized Clinical Trial. Clinical Journal of Pain March/April. 2003;19:87-96.

[99] Huibers MJH, Beurskens AJHM, Van Schayck CP, Bazelmans E, Metsemakers JFM, Knottnerus JA, et al. Efficacy of cognitive-behavioural therapy by general practitioners for unexplained fatigue among employees: Randomised controlled trial. British Journal of Psychiatry March. 2004;184:240-6.

[100] Jalon GGE, Lennon S, Peoples L, Murphy S, Lowe-Strong A. Energy conservation for fatigue management in multiple sclerosis: a pilot randomized controlled trial. Clinical Rehabilitation. 2013;27:63-74.

[101] Jellema P, van der Windt DA, van der Horst HE, Twisk JW, Stalman WA, Bouter LM. Should treatment of (sub)acute low back pain be aimed at psychosocial prognostic factors? Cluster randomised clinical trial in general practice. BMJ. 2005;331:84.

[102] Johnson S, Thornicroft G, Afuwape S, Lesse M, Hughes E, Waingarante S, et al. Effects of training community staff in interventions for substance misue in dual diagnosis patients with psychosis (COMO study): Cluster randomised trial. British Journal of Psychiatry November. 2007;191:451-2.

[103] Jones C, Skirrow P, Griffiths RD, Humphris GH, Ingleby S, Eddleston J, et al. Rehabilitation after critical illness: A randomized, controlled trial. Critical Care Medicine. 2003;31:2456-61.

[104] Joosten EA, de Jong CAJ, de Weert-van Oene GH, Sensky T, van der Staak CPF. Shared Decision-Making Reduces Drug Use and Psychiatric Severity in Substance-Dependent Patients. Psychotherapy & Psychosomatics. 2009;78:245-53.
[105] Jordans MJD, Komproe IH, Tol WA, Kohrt BA, Luitel NP, Macy RD, et al. Evaluation of a classroom-based psychosocial intervention in conflict-affected Nepal: a cluster randomized controlled trial. Journal of Child Psychology & Psychiatry. 2010;51:818-26.
[106] Kaale A, Smith L, Sponheim E. A randomized controlled trial of preschool-based joint attention intervention for children with autism. Journal of Child Psychology & Psychiatry. 2012;53:97-105.
[107] Katon WJ, Lin EH, Von Korff M, Ciechanowski P, Ludman EJ, Young B, et al. Collaborative Care for Patients with Depression and Chronic Illnesses. New England Journal of Medicine. 2010;363:2611-20.
[108] Katon WJ, Von Korff M, Lin EH, Simon G, Ludman E, Russo J, et al. The Pathways Study: A Randomized Trial of Collaborative Care in Patients With Diabetes and Depression. Archives of General Psychiatry. 2004;61:1042-9.
[109] Kehoe CE, Havighurst SS, Harley AE. Tuning in to Teens: Improving Parent Emotion Socialization to Reduce Youth Internalizing Difficulties. Social Development. 2014;23:413-31.
[110] Kellett S, Wilbram M, Davis C, Hardy G. Team consultancy using cognitive analytic therapy: a controlled study in assertive outreach. Journal of Psychiatric & Mental Health Nursing. 2014;21:687-97.
[111] Khumalo-Sakutukwa G, Morin SF, Fritz K, Charlebois ED, van Rooyen H, Chingono A, et al. Project Accept (HPTN 043): A Community-Based Intervention to Reduce HIV Incidence in Populations at Risk for HIV in Sub-Saharan Africa and Thailand. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2008;49:422-31.
[112] Kikkenborg Berg S, Stoier L, Moons P, Zwisler AD, Winkel P, Ulrich Pedersen P. Emotions and health: findings from a randomized clinical trial on psychoeducational nursing to patients with implantable cardioverter defibrillator. The Journal of cardiovascular nursing. 2015;30:197-204.
[113] Klein RG, Abikoff H, Hechtman L, Weiss G. Design and Rationale of Controlled Study of Long-Term Methylphenidate and Multimodal Psychosocial Treatment in Children With ADHD. Journal of the American Academy of Child & Adolescent Psychiatry. 2004;43:792-801.
[114] Kohler L, Meinke-Franze C, Hein J, Fendrich K, Heymann R, Thyrian JR, et al. Does

an Interdisciplinary Network Improve Dementia Care? Results from the IDemUck-Study. Current Alzheimer Research. 2014;11:538-48.

[115] Kronish IM, Rieckmann N, Burg MM, Edmondson D, Schwartz JE, Davidson KW. The effect of enhanced depression care on adherence to risk-reducing behaviors after acute coronary syndromes: Findings from the COPES trial. American Heart Journal. 2012;164:524-9.

[116] Kypri K, McCambridge J, Vater T, Bowe SJ, Saunders JB, Cunningham JA, et al. Web-based alcohol intervention for Maori university students: double-blind, multi-site randomized controlled trial. Addiction. 2013;108:331-8.

[117] Lam CLK, Fong DYT, Chin W-Y, Lee PWH, Lam ETP, Lo YYC. Brief problem-solving treatment in primary care (pst-pc) was not more effective than placebo for elderly patients screened positive of psychological problems. International Journal of Geriatric Psychiatry. 2010;25:968-80.

[118] Lamers F, Jonkers CCM, Bosma H, Kempen GIJM, Meijer JAMJ, Penninx BWJH, et al. A Minimal Psychological Intervention in Chronically Ill Elderly Patients with Depression: A Randomized Trial. Psychotherapy & Psychosomatics. 2010;79:217-26.

[119] Lanken PN, Novack DH, Daetwyler C, Gallop R, Landis JR, Lapin J, et al. Efficacy of an Internet-Based Learning Module and Small-Group Debriefing on Trainees' Attitudes and Communication Skills Toward Patients With Substance Use Disorders: Results of a Cluster Randomized Controlled Trial. Academic Medicine. 2015;90:345-54.

[120] Lash SJ, Stephens RS, Burden JL, Grambow SC, DeMarce JM, Jones ME, et al. Contracting, Prompting, and Reinforcing Substance Use Disorder Continuing Care: A Randomized Clinical Trial. Psychology of Addictive Behaviors. 2007;21:387-97.

[121] Lee KA, Gay CL. Can modifications to the bedroom environment improve the sleep of new parents? Two randomized controlled trials. Research in Nursing & Health. 2011;34:7-19.

[122] Leeuw M, Goossens ME, van Breukelen GJ, de Jong JR, Heuts PH, Smeets RJ, et al. Exposure in vivo versus operant graded activity in chronic low back pain patients: Results of a randomized controlled trial. Pain. 2008;138:192-207.

[123] L'Engle KL, Mwarogo P, Kingola N, Sinkele W, Weiner DH. A Randomized Controlled Trial of a Brief Intervention to Reduce Alcohol Use Among Female Sex Workers in Mombasa, Kenya. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2014;67:446-53.

[124] Leon L, Jover JA, Candelas G, Lajas C, Vadillo C, Blanco M, et al. Effectiveness of an early cognitive-behavioral treatment in patients with work disability due to muscu-

loskeletal disorders. Arthritis & Rheumatism. 2009;61:996-1003.
[125] Lewis FM, Brandt PA, Cochrane BB, Griffith KA, Grant M, Haase JE, et al. The Enhancing Connections Program: A Six-State Randomized Clinical Trial of a Cancer Parenting Program. Journal of Consulting & Clinical Psychology. 2015;83:12-23.
[126] Li L, Hien NT, Lin C, Tuan NA, Tuan LA, Farmer SC, et al. An Intervention to Improve Mental Health and Family Well-Being of Injecting Drug Users and Family Members in Vietnam. Psychology of Addictive Behaviors. 2014;28:607-13.
[127] Limm H, Gundel H, Heinmuller M, Marten-Mittag B, Nater UM, Siegrist J, et al. Stress management interventions in the workplace improve stress reactivity: a randomised controlled trial. Occupational & Environmental Medicine. 2011;68:126-33.
[128] Linares LO, Montalto D, Li M, Oza VS. A Promising Parenting Intervention in Foster Care. Journal of Consulting & Clinical Psychology. 2006;74:32-41.
[129] Lincoln NB, Flannaghan T. Cognitive Behavioral Psychotherapy for Depression Following Stroke: A Randomized Controlled Trial. Stroke. 2003;34:111-5.
[130] Lipsitz JD, Gur M, Vermes D, Petkova E, Cheng J, Miller N, et al. A randomized trial of interpersonal therapy versus supportive therapy for social anxiety disorder. Depression and Anxiety. 2008;25:542-53.
[131] Littbrand H, Lundin-Olsson L, Gustafson Y, Rosendahl E. The Effect of a High-Intensity Functional Exercise Program on Activities of Daily Living: A Randomized Controlled Trial in Residential Care Facilities. Journal of the American Geriatrics Society. 2009;57:1741-9.
[132] Lobban F, Taylor L, Chandler C, Tyler E, Kinderman P, Kolamunnage-Dona R, et al. Enhanced relapse prevention for bipolar disorder by community mental health teams: cluster feasibility randomised trial. British Journal of Psychiatry. 2010;196:59-63.
[133] Lokk JCT, Arnetz BB. Impact of Management Change and an Intervention Program on Health Care Personnel. Psychotherapy & Psychosomatics. 2000;69:79-85.
[134] Lourenco LBdA, Rodrigues RCM, Ciol MA, Sao-Joao TM, Cornelio ME, Dantas RA, et al. A randomized controlled trial of the effectiveness of planning strategies in the adherence to medication for coronary artery disease. Journal of Advanced Nursing. 2014;70:1616-28.
[135] Lu D-F, Hart LK, Lutgendorf SK, Oh H, Schilling M. Slowing progression of early stages of AD with alternative therapies: A feasibility study. Geriatric Nursing. 2013;34:457-64.
[136] Malow RM, Stein JA, McMahon RC, Devieux JG, Rosenberg R, Jean-Gilles M.

Effects of a Culturally Adapted HIV Prevention Intervention in Haitian Youth. Journal of the Association of Nurses in AIDS Care March/April. 2009;20:110-21.
[137] Marshall M, Lockwood A, Green G, Zajac-Roles G, Roberts C, Harrison G. Systematic assessments of need and care planning in severe mental illness: Cluster randomised controlled trial. British Journal of Psychiatry August. 2004;185:163-8.
[138] Martens MP, Smith AE, Murphy JG. The Efficacy of Single-Component Brief Motivational Interventions Among at-Risk College Drinkers. Journal of Consulting & Clinical Psychology. 2013;81:691-701.
[139] Martinsen M, Bahr R, Borresen R, Holme I, Pensgaard AM, Sundgot-Borgen J. Preventing Eating Disorders among Young Elite Athletes: A Randomized Controlled Trial. Medicine & Science in Sports & Exercise. 2014;46:435-47.
[140] Masia Warner C, Fisher PH, Shrout PE, Rathor S, Klein RG. Treating adolescents with social anxiety disorder in school: an attention control trial. Journal of Child Psychology & Psychiatry. 2007;48:676-86.
[141] McCambridge J, Bendtsen M, Karlsson N, White IR, Nilsen P, Bendtsen P. Alcohol assessment and feedback by email for university students: main findings from a randomised controlled trial. British Journal of Psychiatry. 2013;203:334-40.
[142] McCambridge J, Strang J. The efficacy of single-session motivational interviewing in reducing drug consumption and perceptions of drug-related risk and harm among young people: results from a multi-site cluster randomized trial. Addiction. 2004;99:39-52.
[143] McLaughlin TJ, Aupont O, Bambauer KZ, Stone P, Mullan MG, Colagiovanni J, et al. Improving Psychologic Adjustment to Chronic Illness in Cardiac Patients: The Role of Depression and Anxiety. Journal of General Internal Medicine. 2005;20:1084-90.
[144] McSweeney K, Jeffreys A, Griffith J, Plakiotis C, Kharsas R, O'Connor DW. Specialist mental health consultation for depression in Australian aged care residents with dementia: a cluster randomized trial. International Journal of Geriatric Psychiatry. 2012;27:1163-71.
[145] Mealer M, Conrad D, Evans J, Jooste K, Solyntjes J, Rothbaum B, et al. Feasibility and Acceptability of a Resilience Training Program for Intensive Care Unit Nurses. American Journal of Critical Care. 2014;23:e97-e105.
[146] van Meijel B, Kruitwagen C, van der Gaag M, Kahn RS, Grypdonck MHF. An Intervention Study to Prevent Relapse in Patients With Schizophrenia. Journal of Nursing Scholarship. 2006;38:42-9.
[147] Melville JL, Reed SD, Russo J, Croicu CA, Ludman E, LaRocco-Cockburn A, et al.

Improving Care for Depression in Obstetrics and Gynecology: A Randomized Controlled Trial. Obstetrics & Gynecology. 2014;123:1237-46.

[148] Meredith LS, Jackson-Triche M, Duan N, Rubenstein LV, Camp P, Wells KB. Quality Improvement for Depression Enhances Long-term Treatment Knowledge for Primary Care Clinicians. Journal of General Internal Medicine. 2000;15:868-77.

[149] Merritt RK, Price JR, Mollison J, Geddes JR. A cluster randomized controlled trial to assess the effectiveness of an intervention to educate students about depression. Psychological Medicine. 2007;37:363-72.

[150] The Metropolitan Area Child Study Research G, Tolan P. A Cognitive-Ecological Approach to Preventing Aggression in Urban Settings: Initial Outcomes for High-Risk Children. Journal of Consulting & Clinical Psychology. 2002;70:179-94.

[151] Midtgaard J, Christensen JF, Tolver A, Jones LW, Uth J, Rasmussen B, et al. Efficacy of multimodal exercise-based rehabilitation on physical activity, cardiorespiratory fitness, and patient-reported outcomes in cancer survivors: a randomized, controlled trial. Annals of Oncology. 2013;24:2267-73.

[152] Mills M, Loney P, Jamieson E, Gafni A, Browne G, Bell B, et al. A primary care cardiovascular risk reduction clinic in Canada was more effective and no more expensive than usual on-demand primary care - a randomised controlled trial. Health & Social Care in the Community. 2010;18:30-40.

[153] Moadel AB, Bernstein SL, Mermelstein RJ, Arnsten JH, Dolce EH, Shuter J. A Randomized Controlled Trial of a Tailored Group Smoking Cessation Intervention for HIV-Infected Smokers. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2012;61:208-15.

[154] Moffett JAK, Jackson DA, Richmond S, Hahn S, Coulton S, Farrin A, et al. Randomised trial of a brief physiotherapy intervention compared with usual physiotherapy for neck pain patients: outcomes and patients' preference. BMJ. 2005;330:75.

[155] Mohr DC, Carmody T, Erickson L, Jin L, Leader J. Telephone-Administered Cognitive Behavioral Therapy for Veterans Served by Community-Based Outpatient Clinics. Journal of Consulting & Clinical Psychology. 2011;79:261-5.

[156] Monti PM, Rohsenow DJ, Swift RM, Gulliver SB, Colby SM, Mueller TI, et al. Naltrexone and Cue Exposure With Coping and Communication Skills Training for Alcoholics: Treatment Process and 1-Year Outcomes. Alcoholism: Clinical & Experimental Research. 2001;25:1634-47.

[157] Moorey S, Cort E, Kapari M, Monroe B, Hansford P, Mannix K, et al. A cluster

randomized controlled trial of cognitive behaviour therapy for common mental disorders in patients with advanced cancer. Psychological Medicine. 2009;39:713-23.
[158] Morey B, Walker R, Davenport A. More dietetic time, better outcome? Nephron Clin Pract. 2008;109:C173-C80.
[159] Morrell CJ, Slade P, Warner R, Paley G, Dixon S, Walters SJ, et al. Clinical effectiveness of health visitor training in psychologically informed approaches for depression in postnatal women: pragmatic cluster randomised trial in primary care. BMJ. 2009;338:276-85.
[160] Morriss R, Dowrick C, Salmon P, Peters S, Dunn G, Rogers A, et al. Cluster randomised controlled trial of training practices in reattribution for medically unexplained symptoms. The British Journal of Psychiatry. 2007;191:536-42.
[161] Muntingh A, van der Feltz-Cornelis C, van Marwijk H, Spinhoven P, Assendelft W, de Waal M, et al. Effectiveness of Collaborative Stepped Care for Anxiety Disorders in Primary Care: A Pragmatic Cluster Randomised Controlled Trial. Psychotherapy & Psychosomatics. 2013;83:37-44.
[162] Murphy AW, Cupples ME, Smith SM, Byrne M, Byrne MC, Newell J, et al. Effect of tailored practice and patient care plans on secondary prevention of heart disease in general practice: cluster randomised controlled trial. BMJ October. 2009;31.
[163] Naylor MD, Hirschman KB, Hanlon AL, Bowles KH, Bradway C, McCauley KM, et al. Comparison of evidence-based interventions on outcomes of hospitalized, cognitively impaired older adults. Journal of Comparative Effectiveness Research. 2014;3:245-57.
[164] Nourhashemi F, Andrieu S, Gillette-Guyonnet S, Giraudeau B, Cantet C, Coley N, et al. Effectiveness of a specific care plan in patients with Alzheimer's disease: cluster randomised trial (PLASA study). BMJ June. 2010;5.
[165] Perkins KA, Marcus MD, Levine MD, D'Amico D, Miller A, Broge M, et al. Cognitive-Behavioral Therapy to Reduce Weight Concerns Improves Smoking Cessation Outcome in Weight-Concerned Women. Journal of Consulting & Clinical Psychology. 2001;69:604-13.
[166] Peterson MA, Hamilton EB, Russell AD. Starting well: Facilitating the middle school transition. Journal of Applied School Psychology. 2009;25:286-304.
[167] Pfiffner LJ, Hinshaw SP, Owens E, Zalecki C, Kaiser NM, Villodas M, et al. A Two-Site Randomized Clinical Trial of Integrated Psychosocial Treatment for ADHD-Inattentive Type. Journal of Consulting & Clinical Psychology. 2014;82:1115-27.
[168] Pfiffner LJ, Yee Mikami A, Huang-Pollock C, Easterlin B, Zalecki C, McBurnett K. A Randomized, Controlled Trial of Integrated Home-School Behavioral Treatment for

ADHD, Predominantly Inattentive Type. Journal of the American Academy of Child & Adolescent Psychiatry. 2007;46:1041-50.

[169] Phillips G, Bottomley C, Schmidt E, Tobi P, Lais S, Yu G, et al. Well London Phase-1: results among adults of a cluster-randomised trial of a community engagement approach to improving health behaviours and mental well-being in deprived inner-city neighbourhoods. Journal of Epidemiology & Community Health. 2014;68:606-14.

[170] Powers SW, Kashikar-Zuck SM, Allen JR, LeCates SL, Slater SK, Zafar M, et al. Cognitive Behavioral Therapy Plus Amitriptyline for Chronic Migraine in Children and Adolescents: A Randomized Clinical Trial. JAMA. 2013;310:2622-30.

[171] Puschner B, Schofer D, Knaup C, Becker T. Outcome management in in-patient psychiatric care. Acta Psychiatrica Scandinavica. 2009;120:308-19.

[172] Rabow MW, Dibble SL, Pantilat SZ, McPhee SJ. The Comprehensive Care Team: A Controlled Trial of Outpatient Palliative Medicine Consultation. Archives of Internal Medicine. 2004;164:83-91.

[173] Rebergen DS, Bruinvels DJ, Bezemer PD, van der Beek AJ, van Mechelen W. Guideline-Based Care of Common Mental Disorders by Occupational Physicians (CO-OP study): A Randomized Controlled Trial. Journal of Occupational & Environmental Medicine. 2009;51:305-12.

[174] Redhead K, Bradshaw T, Braynion P, Doyle M. An evaluation of the outcomes of psychosocial intervention training for qualified and unqualified nursing staff working in a low-secure mental health unit. Journal of Psychiatric & Mental Health Nursing. 2011;18:59-66.

[175] Richards DA, Hill JJ, Gask L, Lovell K, Chew-Graham C, Bower P, et al. Clinical effectiveness of collaborative care for depression in UK primary care (CADET): cluster randomised controlled trial. BMJ August. 2013;24.

[176] Richards D, Lovell K, Gilbody S, Gask L, Torgerson D, Barkham M, et al. Collaborative care for depression in UK primary care: A randomized controlled trial. Psychological Medicine. 2008;38:279-87.

[177] Robinson LA, Vander Weg MW, Riedel BW, Klesges RC, McLain-Allen B. "Start to stop": results of a randomised controlled trial of a smoking cessation programme for teens. Tobacco Control. 2003;12 Supplement:iv26-iv33.

[178] Rolland Y, Pillard F, Klapouszczak A, Reynish E, Thomas D, Andrieu S, et al. Exercise program for nursing home residents with Alzheimer's disease: A 1-year randomized, controlled trial. Journal of the American Geriatrics Society. 2007;55:158-65.

[179] Rondeau V, Allain H, Bakchine S, Bonet P, Brudon F, Chauplannaz G, et al. General practice-based intervenion for suspecting and detecting dementia in France: A cluster randomized controlled trial. Dementia. 2008;7:433-50.
[180] Ross R, Lam M, Blair SN, Church TS, Godwin M, Hotz SB, et al. Trial of Prevention and Reduction of Obesity Through Active Living in Clinical Settings: A Randomized Controlled Trial. Archives of Internal Medicine. 2012;172:414-24.
[181] Roy-Byrne PP, Craske MG, Stein MB, Sullivan G, Bystritsky A, Katon W, et al. A Randomized Effectiveness Trial of Cognitive-Behavioral Therapy and Medication for Primary Care Panic Disorder. Archives of General Psychiatry. 2005;62:290-8.
[182] Russell AJ, Jassi A, Fullana MA, Mack H, Johnston K, Heyman I, et al. Cognitive behavior therapy for comorbid obsessive-compulsive disorder in high-functioning autism spectrum disorders: A randomized controlled trial. Depression and Anxiety. 2013;30:697-708.
[183] Safren SA, Sprich S, Mimiaga MJ, Surman C, Knouse L, Groves M, et al. Cognitive Behavioral Therapy vs Relaxation With Educational Support for Medication-Treated Adults With ADHD and Persistent Symptoms: A Randomized Controlled Trial. JAMA. 2010;304:875-80.
[184] Saitz R, Cheng D, Winter M, Kim T, Meli S, Allensworth-Davies D, et al. Chronic Care Management for Dependence on Alcohol and Other Drugs: The AHEAD Randomized Trial. JAMA. 2013;310:1156-67.
[185] Samet JH, Raj A, Cheng DM, Blokhina E, Bridden C, Chaisson CE, et al. HERMITAGE-a randomized controlled trial to reduce sexually transmitted infections and HIV risk behaviors among HIV-infected Russian drinkers. Addiction. 2015;110:80-90.
[186] Samuel-Hodge CD, Keyserling TC, Park S, Johnston LF, Gizlice Z, Bangdiwala SI. A Randomized Trial of a Church-Based Diabetes Self-management Program for African Americans With Type 2 Diabetes. Diabetes Educator May/June. 2009;35:439-54.
[187] Sandgren AK, McCaul KD. Short-Term Effects of Telephone Therapy for Breast Cancer Patients. Health Psychology. 2003;22:310-5.
[188] Schiller KR, Luo X, Anderson AJ, Jensen JA, Allen SS, Hatsukami DK. Comparing an Immediate Cessation Versus Reduction Approach to Smokeless Tobacco Cessation. Nicotine & Tobacco Research. 2011;14:902-9.
[189] Schmidt U, Oldershaw A, Jichi F, Sternheim L, Startup H, McIntosh V, et al. Out-patient psychological therapies for adults with anorexia nervosa: randomised controlled trial. British Journal of Psychiatry. 2012;201:392-9.

[190] Schneider JK, Cook JH, Luke DA. Unexpected effects of cognitive-behavioural therapy on self-reported exercise behaviour and functional outcomes in older adults. Age & Ageing. 2011;40:163-8.

[191] Schneider RH, Grim CE, Rainforth MV, Kotchen T, Nidich SI, Gaylord-King C, et al. Stress Reduction in the Secondary Prevention of Cardiovascular Disease: Randomized, Controlled Trial of Transcendental Meditation and Health Education in Blacks. Circulation: Cardiovascular Quality & Outcomes. 2012;5:750-8.

[192] Schonert-Reichl KA, Oberle E, Lawlor MS, Abbott D, Thomson K, Oberlander TF, et al. Enhancing cognitive and social-emotional development through a simple-to-administer mindfulness-based school program for elementary school children: A randomized controlled trial. Developmental Psychology. 2015;51:52-66.

[193] Sells D, Davidson L, Jewell C, Falzer P, Rowe M. The Treatment Relationship in Peer-Based and Regular Case Management for Clients With Severe Mental Illness. Psychiatric Services. 2006;57:1179-84.

[194] Sensky T, Turkington D, Kingdon D, Scott JL, Scott JM, Siddle R, et al. A Randomized Controlled Trial of Cognitive-Behavioral Therapy for Persistent Symptoms in Schizophrenia Resistant to Medication. Archives of General Psychiatry. 2000;57:165-72.

[195] Sharpe H, Schober I, Treasure J, Schmidt U. Feasibility, acceptability and efficacy of a school-based prevention programme for eating disorders: cluster randomised controlled trial. British Journal of Psychiatry. 2013;203:428-35.

[196] Shellman JM, Mokel M, Hewitt N. The Effects of Integrative Reminiscence on Depressive Symptoms in Older African Americans. Western Journal of Nursing Research. 2009;31:772-86.

[197] Shemilt I, Harvey I, Shepstone L, Swift L, Reading R, Mugford M, et al. A national evaluation of school breakfast clubs: evidence from a cluster randomized controlled trial and an observational analysis. Child: Care, Health & Development. 2004;30:413-27.

[198] Shernoff ES, Kratochwill TR. Transporting an Evidence-Based Classroom Management Program for Preschoolers With Disruptive Behavior Problems to a School: An Analysis of Implementation, Outcomes, and Contextual Variables. School Psychology Quarterly. 2007;22:449-72.

[199] Simon GE, Ludman EJ, Unutzer J, Bauer MS, Operskalski B, Rutter C. Randomized trial of a population-based care program for people with bipolar disorder. Psychological Medicine. 2005;35:13-24.

[200] Slade M, McCrone P, Kuipers E, Leese M, Cahill S, Parabiaghi A, et al. Use of

standardised outcome measures in adult mental health services: Randomised controlled trial. British Journal of Psychiatry. 2006;189:330-6.
[201] Speca M, Carlson LE, Goodey E, Angen M. A Randomized, Wait-List Controlled Clinical Trial: The Effect of a Mindfulness Meditation-Based Stress Reduction Program on Mood and Symptoms of Stress in Cancer Outpatients. Psychosomatic Medicine September/October. 2000;62:613-22.
[202] Sripada BN, Henry DB, Jobe TH, Winer JA, Schoeny ME, Gibbons RD. A randomized controlled trial of a feedback method for improving empathic accuracy in psychotherapy. Psychology & Psychotherapy: Theory, Research & Practice. 2011;84:113-27.
[203] Stallard P, Sayal K, Phillips R, Taylor JA, Spears M, Anderson R, et al. Classroom based cognitive behavioural therapy in reducing symptoms of depression in high risk adolescents: pragmatic cluster randomised controlled trial. BMJ October. 2012;6.
[204] Stein BD, Jaycox LH, Kataoka SH, Wong M, Tu W, Elliott MN, et al. A Mental Health Intervention for Schoolchildren Exposed to Violence: A Randomized Controlled Trial. JAMA. 2003;290:603-11.
[205] Steinglass JE, Albano AM, Simpson BH, Wang YP, Zou J, Attia E, et al. Confronting fear using exposure and response prevention for anorexia nervosa: A randomized controlled pilot study. International Journal Of Eating Disorders. 2014;47:174-80.
[206] Stevens VJ, Glasgow RE, Hollis JF, Mount K. Implementation and Effectiveness of a Brief Smoking-Cessation Intervention for Hospital Patients. Medical Care. 2000;38:451-9.
[207] Stewart-Brown S, Patterson J, Mockford C, Barlow J, Klimes I, Pyper C. Impact of a general practice based group parenting programme: quantitative and qualitative results from a controlled trial at 12 months. Archives of Disease in Childhood. 2004;89:519-25.
[208] Stuifbergen AK, Blozis SA, Becker H, Phillips L, Timmerman G, Kullberg V, et al. A randomized controlled trial of a wellness intervention for women with fibromyalgia syndrome. Clinical Rehabilitation. 2010;24:305-18.
[209] Tate DF, Jackvony EH, Wing RR. Effects of Internet Behavioral Counseling on Weight Loss in Adults at Risk for Type 2 Diabetes: A Randomized Trial. JAMA. 2003;289:1833-6.
[210] Thompson EA, Eggert LL, Randell BP, Pike KC. Evaluation of Indicated Suicide Risk Prevention Approaches for Potential High School Dropouts. American Journal of Public Health. 2001;91:742-52.
[211] Thorn BE, Day MA, Burns J, Kuhajda MC, Gaskins SW, Sweeney K, et al. Randomized trial of group cognitive behavioral therapy compared with a pain education control

for low-literacy rural people with chronic pain. Pain. 2011;152:2710-20.
[212] Tiwari A, Leung WC, Leung TW, Humphreys J, Parker B, Ho PC. A randomised controlled trial of empowerment training for Chinese abused pregnant women in Hong Kong. BJOG: An International Journal of Obstetrics & Gynaecology. 2005;112:1249-56.
[213] Tol WA, Komproe IH, Susanty D, Jordans M, Macy RD, De Jong JT. School-Based Mental Health Intervention for Children Affected by Political Violence in Indonesia: A Cluster Randomized Trial. JAMA. 2008;300:655-62.
[214] Tolan P, Gorman-Smith D, Henry D. Supporting Families in a High-Risk Setting: Proximal Effects of the SAFEChildren Preventive Intervention. Journal of Consulting & Clinical Psychology. 2004;72:855-69.
[215] Tross S, Campbell A, Cohen LR, Calsyn D, Pavlicova M, Miele GM, et al. Effectiveness of HIV/STD Sexual Risk Reduction Groups for Women in Substance Abuse Treatment Programs: Results of NIDA Clinical Trials Network Trial. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2008;48:581-9.
[216] Tucker T, Fry CL, Lintzeris N, Baldwin S, Ritter A, Donath S, et al. Randomized controlled trial of a brief behavioural intervention for reducing hepatitis C virus risk practices among injecting drug users. Addiction. 2004;99:1157-66.
[217] Unutzer J, Katon W, Callahan CM, Williams JWJ, Hunkeler E, Harpole L, et al. Collaborative Care Management of Late-Life Depression in the Primary Care Setting: A Randomized Controlled Trial. JAMA. 2002;288:2836-45.
[218] Verbiest MEA, Crone MR, Scharloo M, Chavannes NH, van der Meer V, Kaptein AA, et al. One-Hour Training for General Practitioners in Reducing the Implementation Gap of Smoking Cessation Care: A Cluster-Randomized Controlled Trial. Nicotine & Tobacco Research. 2014;16:1-10.
[219] Vidal R, Bosch R, Nogueira M, Gomez-Barros N, Valero S, Palomar G, et al. Psychoeducation for Adults With Attention Deficit Hyperactivity Disorder vs. Cognitive Behavioral Group Therapy: A Randomized Controlled Pilot Study. Journal of Nervous & Mental Disease. 2013;201:894-900.
[220] Vlasveld MC, van der Feltz-Cornelis CM, Ader HJ, Anema JR, Hoedeman R, van Mechelen W, et al. Collaborative care for sick-listed workers with major depressive disorder: a randomised controlled trial from the Netherlands Depression Initiative aimed at return to work and depressive symptoms. Occupational & Environmental Medicine. 2013;70:223-30.
[221] Waghorn G, Dias S, Gladman B, Harris M, Saha S. A multi-site randomised control-

led trial of evidence-based supported employment for adults with severe and persistent mental illness. Australian Occupational Therapy Journal. 2014;61:424-36.
[222] Wagner GJ, Kanouse DE, Golinelli D, Miller LG, Daar ES, Witt MD, et al. Cognitive-behavioral intervention to enhance adherence to antiretroviral therapy: a randomized controlled trial (CCTG 578). AIDS. 2006;20:1295-302.
[223] Walpole B, Dettmer E, Morrongiello BA, McCrindle BW, Hamilton J. Motivational Interviewing to Enhance Self-Efficacy and Promote Weight Loss in Overweight and Obese Adolescents: A Randomized Controlled Trial. Journal of Pediatric Psychology. 2013;38:944-53.
[224] Webb MS, de Ybarra DR, Baker EA, Reis IM, Carey MP. Cognitive-Behavioral Therapy to Promote Smoking Cessation Among African American Smokers: A Randomized Clinical Trial. Journal of Consulting & Clinical Psychology. 2010;78:24-33.
[225] Wells DL, Dawson P, Sidani S, Craig D, Pringle D. Effects of an Abilities-Focused Program of Morning Care on Residents Who Have Dementia and On Caregivers. Journal of the American Geriatrics Society. 2000;48:442-9.
[226] Wells KB, Sherbourne C, Schoenbaum M, Duan N, Meredith L, Unutzer J, et al. Impact of disseminating quality improvement programs for depression in managed primary care - A randomized controlled trial. Jama-J Am Med Assoc. 2000;283:212-20.
[227] Werch CE, Bian H, Carlson JM, Moore MJ, DiClemente CC, Huang IC, et al. Brief integrative multiple behavior intervention effects and mediators for adolescents. Journal of Behavioral Medicine. 2011;34:3-12.
[228] Wild B, Friederich HC, Gross G, Teufel M, Herzog W, Giel KE, et al. The ANTOP study: focal psychodynamic psychotherapy, cognitive-behavioural therapy, and treatment-as-usual in outpatients with anorexia nervosa–a randomized controlled trial. Trials [Electronic Resource]. 2009;10:23.
[229] Williams JW, Katon W, Lin EH, Noel PH, Worchel J, Cornell J, et al. The Effectiveness of Depression Care Management on Diabetes-Related Outcomes in Older Patients. Annals of Internal Medicine. 2004;140:1015-24.
[230] Wingood GM, DiClemente RJ, Robinson-Simpson L, Lang DL, Caliendo A, Hardin JW. Efficacy of an HIV Intervention in Reducing High-Risk Human Papillomavirus, Nonviral Sexually Transmitted Infections, and Concurrency Among African American Women: A Randomized-Controlled Trial. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2013;63 Supplement:S36-S43.
[231] Wood JJ, Piacentini JC, Southam-Gerow M, Chu BC, Sigman M. Family Cognitive

Behavioral Therapy for Child Anxiety Disorders. Journal of the American Academy of Child & Adolescent Psychiatry. 2006;45:314-21.

[232] Wu L-M, Chiou S-S, Sheen J-M, Lin P-C, Liao YM, Chen H-M, et al. Evaluating the acceptability and efficacy of a psycho-educational intervention for coping and symptom management by children with cancer: a randomized controlled study. Journal of Advanced Nursing. 2014;70:1653-62.

[233] Yoder P, Stone WL. Randomized Comparison of Two Communication Interventions for Preschoolers With Autism Spectrum Disorders. Journal of Consulting & Clinical Psychology. 2006;74:426-35.

[234] Zhou K, Li X, Yan H, Dang S, Wang D-l. Effects of music therapy on depression and duration of hospital stay of breast cancer patients after radical mastectomy. Chinese Medical Journal (English Edition). 2011;124:2321-7.

Appendix B

Shiny application

The sections of this appendix refer to the decision support tool that was developed using Shiny. This was the subject of Chapter 6.

B.1 "ui.R" script

The R code for the "ui.R" script is given below. This determines the appearance of the Shiny application and the levels of the particular inputs that the user can choose.

```
#Shiny app - simulated dataset with contamination only
library(shiny)

#total_sims_be <- load("data/total_sims_part1a.Rda")
# ui.R

shinyUI(
navbarPage("Decision Support Tool",

## Instructions

tabPanel("Instructions",

fluidRow(
column(10, wellPanel(
helpText(tags$b("Background")),

helpText("This tool compares the efficiency of two design options with
    consistent estimators of efficacy (the effect of treatment receipt on
    outcome) in randomised controlled trials where treatment contamination is
```

```
    expected. Contamination is defined as receipt of active intervention
    within the control arm of a trial. The design options are:
"),

helpText(HTML("<ol type=A>
<li>Cluster randomisation at the level at which contamination is anticipated
    to occur with an estimator of average treatment effect (ATE) that
    accounts for clustering of data,</li>
<li>Individual level randomisation, measurement of treatment receipt, and use
    of a randomisation-based estimator of efficacy.</li>
</ol>")),

helpText("Monte Carlo simulations were used to compare these design options.
    It is assumed that:",
tags$ul(
tags$li("There are only two trial arms: active intervention and control,"),
tags$li("There is no treatment non-compliance (active intervention arm) in
    either design option,"),
tags$li("Participants in the control arm receive either the control
    treatment, full active treatment, or some dose of active treatment,"),
tags$li("Cluster randomisation entirely prevents contamination and does not
    cause bias itself.")
)
),

helpText("The tool is presented in two parts. First, with a binary measure of
    treatment receipt, and second, with a continuous measure of treatment
    receipt. For a binary measure of treatment receipt, the efficacy estimand
    is the effect of treatment within a sub-population of participants who
    would receive treatment when offered it and would receive control when
    offered it (i.e. the complier average causal effect). For a continuous
    measure of treatment receipt, the estimand being used is the effect of
    treatment within a sub-population of participants who would receive the
    maximum dose of treatment when offered it and no dose when offered
    control."
),

br(),

helpText(tags$b("Purpose of the tool")),
```

```
helpText(HTML("The aim of the tool is to provide the ratio of the standard
    errors of estimates of efficacy under the design options: SE<sub>A</sub>
    &divide SE<sub>B</sub>. A ratio of greater than one would imply that the
    variance of the estimator of ATE under design option A is greater than
    that of efficacy under design option B. Or put another way, a ratio of
    greater than one means that the efficacy estimator of design option B is
    more precise."
)),

helpText("The tool provides this ratio at the levels of various parameters,
    as set by the user. For binary treatment receipt, these parameters are
    sample size, size of standardised treatment effects, level of ICC, size
    of clusters, proportion of non-contaminators (this parameter represents
    the path from random treatment allocation to treatment receipt). For
    continuous treatment receipt, parameters are sample size, size of
    treatment effects, ICC, size of clusters, size of dose complier stratum,
    and the magnitude of the response within this stratum (size of difference
    between the counterfactual doses). I define the dose compliers as those
    participants who would receive a greater dose of treatment under offer of
    active intervention compared to control. The tool also plots a
    three-dimensional figure of the surface of equivalence between the two
    design options (i.e. the plane at which the precisions of the two design
    options' estimators are equivalent). It is plotted in three dimensions
    because there are three key variables that drive the relative efficiency
    of design option A compared to option B: the level of the ICC, cluster
    size, and proportion of non-contaminators/dose complier stratum."
),

br(),

helpText(tags$b("How to use the tool")),

helpText("The user must first choose the sample size options (100, 200, 500
    or 1000) and standardised treatment effect sizes (0.2, 0.5 or 0.8). The
    setting of these parameters enables the generation of the 3D surface plot
    on the right-hand side of the screen. If the user opts for a binary
    measure of treatment receipt then he or she needs to determine the ICC
    (0.01, 0.02, 0.05 or 0.1), cluster size (5, 10 or 20), and proportion of
    non-contaminators (0.4, 0.5, 0.6, 0.7, 0.8 or 0.9). This proportion is
    also the proportion of latent compliers, or the population who would
    receive treatment when offered and would not receive it when offered
    control. These three parameters are used to plot a coordinate (red ball)
```

```
    in three-dimensional space in the figure. The user may need to use the
    viewing angle toggles beneath the figure to visualise the position of
    this point in relation to the surface."
),

helpText("If the user decides to use a continuous measure of treatment
    receipt, he or she needs to choose the ICC, cluster size, size of dose
    complier stratum (the proportion of the population who would receive a
    greater dose under offer of treatment compared to control; 0.4, 0.5, 0.6,
    0.7, 0.8 or 0.9), and the magnitude of response within this stratum. This
    final parameter represents the difference in dose of active treatment
    under offer of treatment compared to control for the dose compliers. This
    parameter can be set to one of four levels, which are defined by the
    minimum dose of active treatment under its offer and the maximum dose of
    it under offer of control:",
tags$ul(
tags$li("Full dose compliers: participants receive full dose under offer of
    treatment and nothing under offer of control,"),
tags$li("Strong partial dose compliers: participants receive a dose of
    between 80% and 100% of maximum dose under offer of treatment and between
    0% and 20% under control,"),
tags$li("Moderate strength partial dose compliers: participants receive a
    dose of between 50% and 100% of maximum dose under offer of treatment and
    between 0% and 50% under control,"),
tags$li("Weak partial dose compliers: participants receive a dose of between
    0% and 100% of maximum dose under offers of treatment and control (where
    each participant within this stratum receives a greater dose under offer
    of treatment than control).")
)
)

) #End of wellPanel
) #End of column

) #End of fluidRow
), #End of tabPanel


## Binary measure of treatment receipt ##

tabPanel("Binary measure of treatment receipt",
```

```
fluidRow(
column(5, wellPanel(
helpText("Please enter levels for the following input parameters."),

helpText("Modifying the following two parameters will change the location of
    the surface in the plot:"),

selectInput("n_part1",
label = "Choose sample size",
choices = c("100", "200", "500", "1000"),
selected = "100"),

selectInput("beta_part1",
label = "Choose (standardised) treatment effect size",
choices = c("0.2", "0.5", "0.8"),
selected = "0.2"),

helpText("Modify the following three parameters (as well as the previous two)
    to compare the design options (the trial scenario will then be shown as a
    coordinate on the plot):"),

selectInput("ksi_part1",
label = "Choose intraclass correlation coefficient",
choices = c("0.01", "0.02", "0.05", "0.1"),
selected = "0.01"),

selectInput("k_part1",
label = "Choose cluster size",
choices = c("5", "10", "20"),
selected = "5"),

sliderInput("p1_part1",
label = "Choose proportion of non-contaminators",
min = 0.4,
max = 0.9,
step = 0.1,
value=0.8),

helpText(tags$b("It is recommended that the user try different values for the
    intraclass correlation coefficient, cluster size and proportion of
    compliers in order to judge the sensitivity of the results to these
    parameter choices."))
```

```
) #End of wellPanel
), #End of column

column(7,
helpText("The ratio of SEs (SE(A) / SE(B)) was:"),
textOutput("text_part1"),
textOutput("text2_part1"),
helpText("The surface in the figure below represents the equivalence in
    efficiency between the two design options. Below the surface favours the
    efficiency of design option A; above the surface favours option B. Note
    that you may need to change the viewing angle (using the toggles below)
    in order to judge the location of the coordinate.")
), #End of column
column(7,
imageOutput("image_part1")
), #End of column

column(7,
uiOutput('resetable_input_part1'),
actionButton("reset_input_part1", "Reset inputs")
) #End of column

) #End of fluidRow
), #End of tabPanel



## Continuous measure of treatment receipt ##

tabPanel("Continuous measure of treatment receipt",

fluidRow(
column(5, wellPanel(
helpText("Please enter levels for the following input parameters."),

helpText("Modifying the following two parameters will change the location of
    the surface in the plot:"),

selectInput("n_part2",
label = "Choose sample size",
choices = c("100", "200", "500", "1000"),
selected = "100"),
```

```
selectInput("beta_part2",
label = "Choose (standardised) treatment effect size",
choices = c("0.2", "0.5", "0.8"),
selected = "0.2"),

helpText("Modify the following three parameters (as well as the previous two)
    to compare the design options (the trial scenario will then be shown as a
    coordinate on the plot):"),

selectInput("ksi_part2",
label = "Choose intraclass correlation coefficient",
choices = c("0.01", "0.02", "0.05", "0.1"),
selected = "0.01"),

selectInput("k_part2",
label = "Choose cluster size",
choices = c("5", "10", "20"),
selected = "5"),

sliderInput("p1_part2",
label = "Choose proportion of dose compliers",
min = 0.4,
max = 0.9,
step = 0.1,
value=0.8),

helpText("Note that the proportion of dose compliers is the size of the
    stratum of participants who would receive a greater dose of active
    intervention under offer of treatment than under offer of control."),

selectInput("F_range_part2",
label = "Choose magnitude of response for those participants who would
    receive greater dose under offer of treatment compared to control",
choices = c("Full dose compliers"=0, "Strong partial dose compliers"=2,
"Moderate strength partial dose compliers"=5, "Weak partial dose
    compliers"=10),
selected = 0),

helpText("This parameter refers to the level of the difference in potential
    (counterfactual) dose between the offers of treatment and control. \"Full
    dose compliers\" receive full dose under offer of treatment and nothing
```

```
    under offer of control; \"Strong partial dose compliers\" receive a dose
    of between 80% and 100% of maximum dose under offer of treatment and
    between 0% and 20% under control; \"Moderate strength partial dose
    compliers\" receive a dose of between 50% and 100% of maximum dose under
    offer of treatment and between 0% and 50% under control; \"Weak partial
    dose compliers\" receive a dose of between 0% and 100% of maximum dose
    under offers of treatment and control (where each participant within this
    stratum receives a greater dose under offer of treatment than control)."),

helpText(tags$b("It is recommended that the user try different values for the
    intraclass correlation coefficient, cluster size, proportion of dose
    compliers and magnitude of compliance (within the dose complier stratum)
    in order to judge the sensitivity of the results to these parameter
    choices."))

) #End of wellPanel
), #End of column

column(7,
helpText("The ratio of SEs (SE(A) / SE(B)) was:"),
textOutput("text_part2"),
textOutput("text2_part2"),

br(),

textOutput("help_text_part2"),
tags$head(tags$style("#help_text_part2{color: grey;}"
)
)

#helpText("The surface in the figure below represents the equivalence in
    efficiency between the two design options. Below the surface favours the
    efficiency of design option A; above the surface favours option B. Note
    that you may need to change the viewing angle (using the toggles below)
    in order to judge the location of the coordinate.")
), #End of column

column(7,
#if (input$F_range_part2 < 4) {
imageOutput("image_part2")
), #End of column
```

```
column(7,
uiOutput('resetable_input_part2'),
actionButton("reset_input_part2", "Reset inputs")
) #End of column

) #End of fluidRow
) #End of tabPanel

) #End of narbarPage
) #End of UI
```

B.2 "server.R" script

The R code for the "server.R" script is given below. This renders the functionality of the Shiny application, e.g. passes levels of input parameters that are set by the user to the function that plots the 3D figure.

```
#Shiny app - simulated dataset with contamination only
# server.R
library(shiny)

#Source scripts, loading of data, and loading of libraries goes outside the
    shinyServer function (because this is only run once):
#setwd("U:/PhD/Simulations (primary research question)/Shiny app")
total_sims_be <- load("data/bothparts_SE.Rda")
source("data/print_output_part1.R", local=TRUE)
source("data/3d_plot_working_part1.R", local=TRUE)
source("data/print_output_part2.R", local=TRUE)
source("data/3d_plot_working_part2.R", local=TRUE)

shinyServer(
function(input, output) {

### Part 1: Binary measure of treatment receipt ###

#This gives the 3D plot:
output$image_part1 <- renderPlot({
plot(plot3dfnc_part1(n3_part1=input$n_part1,
```

```
beta3_part1=input$beta_part1,
pt.x2_part1=as.numeric(input$ksi_part1)*100, # Ksi (ICC)
pt.y2_part1=as.numeric(input$k_part1)*0.25, # K
pt.z2_part1=input$p1_part1*5, # IV
screen.x2_part1=input$screen.x_part1,
screen.y2_part1=0.000001,
screen.z2_part1=input$screen.z_part1,
zoom2_part1=input$zoom_part1))}, width=700
) #End of renderPlot

#The next items give the text info at top of right hand panel:
output$text_part1 <- renderText({
print(round(print_output_part1(n2_part1=input$n_part1,
beta2_part1=input$beta_part1,
ksi2_part1=input$ksi_part1,
k2_part1=input$k_part1,
p12_part1=input$p1_part1,
delta2_part1=0.5),3))
}) #End of renderText

output$text2_part1 <- renderText({
if (print_output_part1(n2_part1=input$n_part1, beta2_part1=input$beta_part1,
    ksi2_part1=input$ksi_part1, k2_part1=input$k_part1,
    p12_part1=input$p1_part1, delta2_part1=0.5) < 1) {
paste("The ratio, which is less than one, suggests that design option option
    A is more efficient.")
}

else if (print_output_part1(n2_part1=input$n_part1,
    beta2_part1=input$beta_part1, ksi2_part1=input$ksi_part1,
    k2_part1=input$k_part1, p12_part1=input$p1_part1, delta2_part1=0.5) > 1) {
paste("The ratio, which is more than one, suggests that design option option
    B is more efficient.")
}
}) #End of renderText

#The next items provide the controls for viewing of the 3D graph:
output$resetable_input_part1 <- renderUI({
times <- input$reset_input_part1
div(id=letters[(times %% length(letters)) + 1],
fluidRow(
column(6,
```

```
sliderInput("screen.x_part1",
label = "Choose rotation in vertical plane",
min = -90,
max = 0,
step = 5,
value = -80),
sliderInput("zoom_part1",
label = "Choose zoom",
min = 0,
max = 1,
step = 0.01,
value = 0.74)
), #End of column
column(6,
sliderInput("screen.z_part1",
label = "Choose rotation in horizontal plane",
min = -90,
max = 90,
step = 5,
value = -65)
) #End of column
) #End of fluidRow
) #End of div
}) #End of renderUI


### Part 2: Continuous measure of treatment receipt ###

#This gives the 3D plot:
output$image_part2 <- renderPlot({
if (input$F_range_part2 == 0 | input$F_range_part2 == 2) {
plot(plot3dfnc_part2(n3_part2=input$n_part2,
beta3_part2=input$beta_part2,
F_range3_part2=input$F_range_part2,
pt.x2_part2=as.numeric(input$ksi_part2)*100, # Ksi (ICC)
pt.y2_part2=as.numeric(input$k_part2)*0.25, # K
pt.z2_part2=input$p1_part2*5, # IV
screen.x2_part2=input$screen.x_part2,
screen.y2_part2=0.000001,
screen.z2_part2=input$screen.z_part2,
zoom2_part2=input$zoom_part2))
```

```
} #End of if
}, width=700) #End of renderPlot

#The next items give the text info at top of right hand panel:
output$text_part2 <- renderText({
print(round(print_output_part2(n2_part2=input$n_part2,
beta2_part2=input$beta_part2,
F_range2_part2=input$F_range_part2,
ksi2_part2=input$ksi_part2,
k2_part2=input$k_part2,
p12_part2=input$p1_part2,
delta2_part2=0.5),3))
}) #End of renderText

output$text2_part2 <- renderText({
if (print_output_part2(n2_part2=input$n_part2, beta2_part2=input$beta_part2,
    F_range2_part2=input$F_range_part2, ksi2_part2=input$ksi_part2,
    k2_part2=input$k_part2, p12_part2=input$p1_part2, delta2_part2=0.5) < 1) {
paste("The ratio, which is less than one, suggests that design option option
    A is more efficient.")
}

else if (print_output_part2(n2_part2=input$n_part2,
    beta2_part2=input$beta_part2, F_range2_part2=input$F_range_part2,
    ksi2_part2=input$ksi_part2, k2_part2=input$k_part2,
    p12_part2=input$p1_part2, delta2_part2=0.5) > 1) {
paste("The ratio, which is more than one, suggests that design option option
    B is more efficient.")
}
}) #End of renderText


output$help_text_part2 <- renderText({
if (input$F_range_part2 == 0 | input$F_range_part2 == 2) {
paste("The surface in the figure below represents the equivalence in
    efficiency between the two design options. Below the surface favours the
    efficiency of design option A; above the surface favours option B. Note
    that you may need to change the viewing angle (using the toggles below)
    in order to judge the location of the coordinate.")
}

else if (input$F_range_part2 == 5 | input$F_range_part2 == 10) {
```

```
paste("It is not possible to plot the plane of equivalence given the levels
    of parameters that have been selected. This is because trial design
    option A is more efficient at every simulated level of intraclass
    correlation coefficient, cluster size, proportion of dose compliers, and
    level of magnitude of difference in potential treatment receipt between
    the trial arms for those participants who would receive greater dose
    under offer of treatment compared to control.")
}
}) #End of renderText


#The next items provide the controls for viewing of the 3D graph:
output$resetable_input_part2 <- renderUI({
times <- input$reset_input_part2
div(id=letters[(times %% length(letters)) + 1],
fluidRow(
column(6,
sliderInput("screen.x_part2",
label = "Choose rotation in vertical plane",
min = -90,
max = 0,
step = 5,
value = -80),
sliderInput("zoom_part2",
label = "Choose zoom",
min = 0,
max = 1,
step = 0.01,
value = 0.74)
), #End of column
column(6,
sliderInput("screen.z_part2",
label = "Choose rotation in horizontal plane",
min = -90,
max = 90,
step = 5,
value = -65)
) #End of column
) #End of fluidRow
) #End of div
}) #End of renderUI
```

```
}  #end of function(input, output)
)  #end of shinyServer
```

B.3 Print output script

The R code for a function that prints the efficiency ratio at particular levels specified by the user is given below.

```
# Print SE ratio function:

print_output_part1 <- function(n2_part1, beta2_part1, ksi2_part1, k2_part1,
    p12_part1, delta2_part1) {

load("data/bothparts_SE.Rda")
#load("U:/PhD/Simulations (primary research question)/Shiny
    app/data/bothparts_SE.Rda")

print(subset(bothparts_SE,subset=n_part1==n2_part1 & beta_part1==beta2_part1
    & ksi_part1==ksi2_part1 & k_part1==k2_part1 & p1_part1==p12_part1 &
    delta_part1==delta2_part1)$SE_ratio_part1)

} #End of function
```

B.4 3D plot script

The R code for a function that plots the 3D figure is given below.

```
#Code for creating 3D plot of simulated data with a marked coordinate.

library("misc3d")
library("rgl")

plot3dfnc_part1 <- function(n3_part1, beta3_part1, pt.x2_part1, pt.y2_part1,
    pt.z2_part1, screen.x2_part1, screen.y2_part1=0, screen.z2_part1,
```

```
    zoom2_part1) {
# x: Ksi, ICC (1, 2, 5, 10)
# y: K, cluster size (1.25, 2.5, 5)
# z: Strength of IV (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0)

load("data/bothparts_SE.Rda")


#3D - create an array first:
k <- c(5,10,20)
ksi <- c(0.01,0.02,0.05,0.10)
p1 <- c(0.4,0.5,0.6,0.7,0.8,0.9)
array111 <- array(subset(bothparts_SE,subset = n_part1==n3_part1 &
    beta_part1==beta3_part1 & delta_part1==0.5)$SE_ratio_part1,
c(4,3,6),
dimnames=list(c(0.01,0.02,0.05,0.1),c(5,10,20),c(0.4,0.5,0.6,0.7,0.8,0.9))) #
    4 rows of ksi, 3 columns of k, 6 tables of p1
array111

#3D Plot of above:
xlim <- c(1, 10) # ksi
ylim <- c(1.25, 5) # cluster size
zlim <- c(0, 5) # strength of IV
v111 <- contour3d(array111, x=ksi*100, y=k/4, z=p1*5, 1, color = "lightblue",
    color2 = "lightblue", alpha=0.7, draw = FALSE, add=TRUE) # alpha is
    transparency
library("lattice")
#pt.x <- 10   # Ksi, ICC (1, 2, 5, 10)
#pt.y <- 2.5  # K, cluster size (1.25, 2.5, 5)
#pt.z <- 4.5    # Strength of IV (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0,
    4.5, 5.0)
w111 <- wireframe(matrix(zlim[1], 2, 2) ~ rep(xlim, 2) * rep(ylim, each = 2),
xlim = xlim, ylim = ylim, zlim = zlim,
aspect = c(diff(ylim) / diff(xlim), diff(zlim) / diff(xlim)),
xlab = "Intraclass \n correlation coefficient", ylab = "Cluster size",
zlab = "Strength \n of IV",
scales = list(arrows = FALSE, col = "black",
x = list(at=c(1,2,5,10),
labels=c("0.01","0.02","0.05","0.10") ), # ksi (0.01, 0.02, 0.05, 0.10)
y = list(at=c(1.25,2.5,5),
labels=c("5","10","20") ), # k (5, 10, 20)
z = list(at=c(0,0.5,1,1.5,2,2.5,3,3.5,4,4.5,5),
```

```
labels=c("0","0.1","0.2","0.3","0.4","0.5","0.6","0.7","0.8","0.9","1.0"))),
    # strength of IV
screen = list(z = screen.z2_part1, x = screen.x2_part1, y = screen.y2_part1),
zoom=zoom2_part1,
par.settings = list(axis.line = list(col = "transparent")),
panel.3d.wireframe = function(x, y, z, rot.mat, distance,
xlim.scaled, ylim.scaled,
zlim.scaled, ...) {
panel.3dscatter(x=pt.x2_part1, y=pt.y2_part1, z=pt.z2_part1, rot.mat,
    distance=distance,
xlim.scaled=xlim.scaled, ylim.scaled=ylim.scaled,
zlim.scaled=zlim.scaled, type="p", col=2, # col=2 gives the point the colour
    red
cex=3, pch=16, .scale=TRUE, ...)
scale <- c(diff(xlim.scaled) / diff(xlim),
diff(ylim.scaled) / diff(ylim),
diff(zlim.scaled) / diff(zlim))
shift <- c(mean(xlim.scaled) - mean(xlim) * scale[1],
mean(ylim.scaled) - mean(ylim) * scale[2],
mean(zlim.scaled) - mean(zlim) * scale[3])
P <- rbind(cbind(diag(scale), shift), c(0, 0, 0, 1))
rot.mat <- rot.mat %*% P
drawScene(v111, screen = NULL, R.mat = rot.mat,
distance = distance, add = TRUE, scale = FALSE,
light = c(0, 0, 1), engine = "grid")
})

}
```

Appendix C

Application of efficacy estimators to D6

The sections of this appendix refer to the code used for applying estimators of CACE to the D6 trial data.

C.1 Stata code for estimator E-IV6 (Bloom/ratio) with bootstrap standard error

The following Stata code was used for the main analysis of efficacy with binary treatment receipt. The code represents a Stata bootstrap program for estimating the Bloom/ratio estimator (**E-IV6**) at 18 months after randomisation with sampling at the level of the nurse.

```
program iv_est_bloom_iv6, rclass
   regress trt_rec2 trial_arm i.ethnicity_new i.education
   local trial_arm_alpha = _b[trial_arm]
   regress HbA1c9 trial_arm i.borough i.phase HBA1cm_base i.ethnicity_new i.education
   local trial_arm_beta = _b[trial_arm]
   return scalar iv = `trial_arm_beta'/`trial_arm_alpha'
end
bootstrap r(iv), reps(200) seed(1987) cluster(nurse) : iv_est_bloom_iv6
```

C.2 Stata code for estimator E-IV5 (modified Bloom/ratio) with bootstrap standard error

The following Stata code was used as one of the sensitivity analyses of efficacy with binary treatment receipt. The code represents a Stata bootstrap program for estimating the modified Bloom/ratio estimator (**E-IV5**) at 18 months after randomisation with sampling at the level of the nurse.

```
program iv_est_iv5, rclass
   regress trt_rec2 trial_arm
   local trial_arm = _b[trial_arm]
   local trial_arm_contr = _b[_cons]
   summarize HbA1c18 if trial_arm==0 & trt_rec2==1
   local HbA1c18_01 = r(mean)
   summarize HbA1c18 if trial_arm==0 & trt_rec2==0
   local HbA1c18_00 = r(mean)
   summarize HbA1c18 if trial_arm==1 & trt_rec2==1
   local HbA1c18_11 = r(mean)
   summarize HbA1c18 if trial_arm==1 & trt_rec2==0
   local HbA1c18_10 = r(mean)
   return scalar iv = ((('trial_arm'+'trial_arm_contr')*'HbA1c18_11') +   ///
      ((1-('trial_arm'+'trial_arm_contr'))*'HbA1c18_10') -   ///
      ('trial_arm_contr'*'HbA1c18_01') -   ///
      ((1-'trial_arm_contr')*'HbA1c18_00'))/'trial_arm'
end
bootstrap r(iv), reps(200) seed(1987) cluster(nurse) : iv_est_iv5
```

C.3 MPlus code for STR3 estimator

The following MPlus code was used for one of the sensitivty analyses of efficacy with binary treatment receipt. The code represents a mixture model estimating efficacy (estimator **E-STR3**) at nine months after randomisation and assuming latent ignorability.

```
TITLE:
```

```
  D6 HbA1C - Efficacy analysis;

DATA:
  File is "str3_model_v1.dat" ;

VARIABLE:
  Names are d6no borough gender bl_BMI phase nurse HBA1cm_bl education HbA1c18
    HbA1c15 HbA1c9 resp9 resp15 resp18 trial_arm trt_rec2 noncomp2 contam2
    ethnicity_new
  Classes C(3);
  Categorical contam2 noncomp2 resp9;
  Usevariables trial_arm HbA1c9 contam2 noncomp2
    phase HBA1cm_bl resp9 gender bl_BMI
    b2 b3 b4 b5 ed2 ed3 eth1 eth2;
  Missing are all (99) ;

ANALYSIS:
  Type = Mixture ;

DEFINE:
  b2 = borough == 2;
  b3 = borough == 3;
  b4 = borough == 4;
  b5 = borough == 5;
  ed2 = education == 2;
  ed3 = education == 3;
  eth1 = ethnicity_new == 1;
  eth2 = ethnicity_new == 2;

MODEL:
  %OVERALL%
  HbA1c9 ON trial_arm b2 b3 b4 b5 phase HBA1cm_bl eth1 eth2;
  resp9 ON b2 b3 b4 b5 phase HBA1cm_bl eth1 eth2;
  C#1 ON gender ed2 ed3;
  C#2 ON bl_BMI b2 b3 b4 b5;

  %C#1%        !always takers class
  [contam2$1@-15];
  [noncomp2$1@-15] ;
```

[HbA1c9];

HbA1c9 ON trial_arm@0 ;

resp9 ON trial_arm@0 ;

%C#2% !never takers class

[contam2$1@15] ;

[noncomp2$1@15] ;

[HbA1c9];

HbA1c9 ON trial_arm@0 ;

resp9 ON trial_arm@0 ;

%C#3% !(latent) complier class

[contam2$1@15] ;

[noncomp2$1@-15] ;

[HbA1c9];

HbA1c9 ON trial_arm ;

resp9 ON trial_arm ;